Enhancing Statistician Power: Flexible Covariate-Adjusted Semiparametric Inference for Randomized Studies with Multivariate Outcomes

Advisor: Professor Victor De Gruttola Alisa Stephens

Enhancing Statistician Power: Flexible Covariate-adjusted Semiparametric Inference for Randomized Studies with Multivariate Outcomes

Abstract

It is well known that incorporating auxiliary covariates in the analysis of randomized clinical trials (RCTs) can increase efficiency. Questions still remain regarding how to flexibly incorporate baseline covariates while maintaining valid inference. Recent methodological advances that use semiparametric theory to develop covariate-adjusted inference for RCTs have focused on independent outcomes. In biomedical research, however, cluster randomized trials and longitudinal studies, characterized by correlated responses, are commonly used. We develop methods that flexibly incorporate baseline covariates for efficiency improvement in randomized studies with correlated outcomes.

In Chapter 1, we show how augmented estimators may be used for cluster randomized trials, in which treatments are assigned to groups of individuals. We demonstrate the potential for imbalance correction and efficiency improvement through consideration of both cluster- and individual-level covariates. To improve small-sample estimation, we consider several variance adjustments. We evaluate this approach for continuous and binary outcomes through simulation and apply it to the *Young Citizens* study, a cluster randomized trial of a community behavioral intervention for HIV prevention in Tanzania.

Chapter 2 builds upon the previous chapter by deriving semiparametric locally efficient estimators of marginal mean treatment effects when outcomes are correlated. Estimating equations are determined by the efficient score under a mean model for marginal effects when data contain baseline covariates and exhibit correlation. Locally efficient es-

timators are implemented for longitudinal data with continuous outcomes and clustered data with binary outcomes. Methods are illustrated through application to AIDS Clinical Trial Group Study 398, a longitudinal randomized study that compared various protease inhibitors in HIV-positive subjects.

In Chapter 3, we empirically evaluate several covariate-adjusted tests of intervention effects when baseline covariates are selected adaptively and the number of randomized units is small. We demonstrate that randomization inference preserves type I error under model selection while tests based on asymptotic theory break down. Additionally, we show that covariate adjustment typically increases power, except at extremely small sample sizes using liberal selection procedures. Properties of covariate-adjusted tests are explored for independent and multivariate outcomes. We revisit *Young Citizens* to provide further insight into the performance of various methods in small-sample settings.

Contents

Acknowledgments

"We can only be said to be alive in those moments when our hearts are conscious of our treasures."

-Thornton Wilder

I am incredibly grateful to my advisors, Victor De Gruttola and Eric Tchetgen Tchetgen, for their guidance, mentorship, and confidence in me. I thank Victor for his extraordinary vision and passion, and Eric for challenging me to learn topics I never thought I could. I also thank my committee member Xihong Lin for her helpful feedback and generosity.

I am indebted to my parents, Edward and Elsa Stephens, for their continued love and support, for providing me with the tools I needed to succeed, and for trusting me to choose a career in biostatistics. To my brothers, Marc and Seth, I am especially thankful for the ways in which you encouraged me as a child, and for the countless ways that you have been actively present in every step of my life. I would not be who I am without you.

I am also grateful to my sweetheart, Jonathan Shields, for celebrating my every achievement, and for keeping his promise to buy his plane tickets no matter where I decided to go for graduate school.

Finally, I would like to thank countless friends and loved ones for sharing in my experience. I extend the warmest of thanks to Loni Philip Tabb for being the best Student Buddy ever, Anna Snavely for being my classroom companion, and Christina McIntosh for giving me a reason to set a good example.

I am extremely blessed to have each one of you in my life.

Augmented GEE for Improving Efficiency and Validity of Estimation in Cluster Randomized Trials by Leveraging Cluster- and Individual-level Covariates

Alisa J. Stephens, Eric J. Tchetgen Tchetgen, and Victor De Gruttola

Department of Biostatistics

Harvard University

1.1 Introduction

1.1.1 Traditional and Cluster Randomized Trials

Randomized clinical trials (RCTs) are recognized as the gold standard in medical research for evaluating new treatments. Cluster or group randomized trials (GRTs), which assign treatment to groups of individuals, are advantageous when interaction among subjects within a group may impact their respective outcomes. GRTs are therefore especially relevant for assessing prevention and treatment methods for infectious diseases, where subjects within a geographical unit such as a neighborhood, school, or workplace may infect each other. For example, in vaccine studies, a subject's vaccination status may impact health outcomes not only for that subject but for others as well. Clustered designs also have the advantage of reducing the potential for contamination of effects caused by sharing of information or medication between treated and control subjects. Similarly, group treatment assignment can enhance compliance as subjects within a group are given the same regimen to follow. In some cases, the intervention may be administered at the cluster-level, such as in studies involving schools or medical practices. Klar and Donner provide several examples of intervention trials in which groups were randomized for medical, political, or logistical reasons {Klar and Donner (2000)}.

Although intervention is assigned at the group level, interest often lies in performing inference on the individual. Generally, subjects within a group are expected to be more similar than subjects in different groups, inducing dependence across study subjects. Cluster randomized designs thus present the additional challenge of accounting for correlation among group members. Standard approaches for estimating treatment effects when responses are correlated include maximum likelihood for generalized linear mixed models (GLMM) and generalized estimating equations (GEE) for restricted mean models {Laird and Ware (1982); Liang and Zeger (1986)}. To estimate the marginal effect in a binary treatment setting, one typically fits a model including an intercept and treatment term. The relevant GLMM is defined by the model $E(Y_{ij}|A_i, b_i) = g(\beta_0 + \beta_{1^c} A_i + b_i)$, where

Y_{ij} denotes the outcome for the j_{th} individual in the i_{th} cluster, A_i is an indicator for treatment, b_i is a random effect inducing correlation among subjects within a cluster, and g($\cdot$) is a monotone link function. The outcome Y_{ij} and random effect b_i are assumed to follow a particular distribution. We note here that β_{1^c} is interpreted as a cluster-specific treatment effect, but marginalizes over all other covariates. In the analogous GEE approach, estimating equations are constructed following the mean model

$$E(Y_{ij}|A_i) = g(A_i;\beta) = g(\beta_0 + \beta_1 A_i), \tag{1.1}$$

where correlation is accounted for by incorporating a working covariance matrix $\mathbf{V_w}$. For cluster randomized designs, independence or exchangeable structure is generally assumed. An advantage of the GEE approach is that consistency of $\hat{\beta}$, the estimate of β, only requires that the mean $g(A_i;\beta)$ is correctly specified, in which case, $\hat{\beta}$ is asymptotically normal for all $\mathbf{V_w}$ and efficient when $\mathbf{V_w}$ takes the true form of $\mathbf{V}$, the variance of response vector $\mathbf{Y}_i$. The exact form of the GEE is reviewed in the following section. GEE differ from maximum likelihood estimation in mixed models by treating correlation as a nuisance parameter. Additionally, GLMM require full specification of the distribution of Y_{ij}, while GEE follow from semiparametric theory and only specify the first moment of Y_{ij} while requiring the second moment to be finite. Unlike GLMM, GEE do not make any assumptions about cluster effects, and thus provide a population-averaged effect estimate in contrast to the GLMM cluster-specific estimate. In either approach, treatment is evaluated through inference on β_1.

A second challenge presented by cluster randomized designs is that the number of available experimental units may be fairly small. Inference for model-based methods relies on asymptotic theory, which may not be applicable in trials with relatively few clusters. For GEE, several studies have shown that the sandwich variance estimator typically underestimates the variability of parameter estimates and consequently results in inference that is too liberal {Gunsolley et al. (1995)}. A number of adjustment methods for small sample analysis have been proposed {Fay and Graubard (2001); Mancl and DeRouen (2001); Pan and Wall (2002); Thornquist and Anderson (1992); Kauermann and Carroll (2001)}. These adjustments generally take one of two strategies; they account for the

variability in the sandwich estimator or correct for its small-sample bias. None of these methods have been uniformly adopted.

The number of available experimental units also affects the degree to which randomization successfully balances baseline characteristics across treatment groups. RCTs with large sample sizes assure a reasonable degree of balance in covariate profiles with high probability, but GRTs often have smaller numbers of experimental units, and therefore provide less assurance of balance Murray et al. (2004). GRTs are also likely to contain subject heterogeneity in cluster-level and individual-level characteristics that can influence estimated treatment effects. Clustered designs therefore require methods that permit controlling for imbalances at the cluster and subject levels.

1.1.2 Methods for Covariate Adjustment in Randomized Trials

Traditionally, adjustment for residual imbalance has been achieved by adding covariates Z_i, X_{ij} to a model for the effect of treatment on some outcome. The adjusted model for Y_{ij} is defined by $E(Y_{ij}|A_i, Z_i, X_{ij}) = g(\beta_0 + \beta_{1^*}A_i + \beta_Z Z_i + \beta_X X_{ij})$, where Z_i is a vector of covariates shared by all subjects within the i_{th} cluster, and X_{ij} is a subject-specific vector of measurements. Standard approaches such as mixed models and GEE can incorporate adjustment at both levels. With the exception of linear and log-linear models, the conditional model differs from the marginal model (1.1) in the interpretation of β_{1^*}. Inference on β_{1^*} is also affected by the presence of baseline covariates. For uncorrelated continuous outcomes and an identity link function relating covariates to the mean, it has been shown that when X and Y are correlated, β_{1^*} is more precise than the unadjusted estimator {Pocock et al. (2002); Tsiatis et al. (2008)}. No direct relationship between the efficiency of β_1 (1.1) and β_{1^*} has been established for non-linear models {Robinson and Jewell (1991)} or correlated outcomes. To provide an alternative that makes fewer parametric assumptions, Gail et al. (1996) proposed a permutation approach to covariate adjustment in GRTs. Parametric models are used for adjustment, and permutation inference is conducted on the cluster-averaged model-based residuals. Permutation tests are

guaranteed to be valid even for small samples, unlike modeling approaches. A similar model-based permutation approach using an optimally weighted combination of residuals was developed by {Braun and Feng (2001)}.

Recent methodological developments in covariate adjustment for RCTs include van der Laan's Targeted Maximum Likelihood {van der Laan and Rubin (2006)} and Tsiatis' augmentation approach {Tsiatis et al. (2008); Zhang et al. (2008)}. These methods adapt semiparametric theory developed by Robins (1999) and Robins et al. (1994) for observational studies with time-varying exposures and missing data problems, respectively. RCTs may be conceptualized theoretically in either framework, with counterfactual outcomes under the treatment not received considered missing, or as observational studies with a known probability of point exposure. Robins et al. (1994) and Robins (1999) characterize the efficient influence function in these settings. van der Laan and Tsiatis solve the set of estimating functions determined by the efficient score using two different approaches, which are equivalent in the absence of model misspecification.

Targeted Maximum Likelihood Estimation (tMLE) is an iterative procedure that involves adding a cleverly defined covariate to standard regression models. Upon convergence, the tMLE estimator solves the efficient influence function for the parameter of interest, resulting in bias reduction and efficiency improvement relative to maximum likelihood. TMLE is currently available for independent binary, continuous, and time-to-event outcomes {Moore and van der Laan (2009b,a)}. More recently it has been extended to correlated data {Tuglus and van der Laan (2010)}. Tsiatis' approach involves directly solving a set of augmented estimating equations determined by the efficient influence function for the marginal treatment effect {Zhang et al. (2008)}. This method has been explored for continuous, binary, and discrete survival outcomes {Leon et al. (2003); Tsiatis et al. (2008); Zhang et al. (2008); Zhang and Gilbert (2010)}. Current applications of the augmentation method have focused on independent outcomes, with the exception of a simulation study based on the linear mixed model {Zhang et al. (2008)}.

While tMLE simultaneously uses baseline covariates from treatment and control

groups to target treatment effect estimation, Tsiatis' method separates covariate adjustment and treatment evaluation. It also has the added advantage of allowing separate adjustment for baseline covariates within treatment arms. If done by separate statistical groups that do not share data, this approach reduces the risk that adjustment models are chosen to yield the most significant result. Even without decoupling of adjustment and treatment effect estimation, all covariate adjustment methods can be made objective by prespecifying the adjustment strategy.

1.2 The Simple Augmented GEE

This section demonstrates the use of augmented estimating equations in analyses of cluster randomized trials. In such a trial, m clusters of size n_i, $i = 1, ..., m$, are randomized to either treatment $(A_i = 1)$ or control $(A_i = 0)$ with probability $P(A_i = 1) = \pi$. To motivate the augmented GEE, we first review Standard GEE. Let Y_{ij} denote the response for the j_{th} individual in the i_{th} cluster. $\mathbf{Y}_i = (Y_{i1},\ Y_{i2}, ..., Y_{in_i})^{\mathrm{T}}$, where n_i is the number of subjects within the i_{th} cluster. GEE for the marginal treatment effect are defined by the mean model (1.1), where β is a p-dimensional parameter. An estimator for β is obtained by solving the estimating equations

$$\sum_{i=1}^{m} \psi_i(\mathbf{Y}, A; \beta) = \sum_{i=1}^{m} \mathbf{D}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \mathbf{g}(A_i; \beta)\} = \mathbf{0}, \tag{1.2}$$

where $\mathbf{D}_i = \frac{\partial \mathbf{g}(A_i;\beta)}{\partial \beta^{\mathrm{T}}}$, $\mathbf{V}_i = V_\phi^{1/2} R\{\alpha(A_i)\} V_\phi^{1/2}$, and bold $\mathbf{g}(A_i; \beta)$ denotes the n_i-dimensional link function for the outcome vector $\mathbf{Y}_i$. Covariance matrix $\mathbf{V}_i$ is determined by the $n_i \times n_i$ matrix function $v(A_i)$. The variance function $v(A_i)$ is a product of the diagonal matrix $\mathbf{V}_\phi$, where $V_{\phi_{i,i}}$ is the variance of Y_{ij}, and correlation matrix $\mathbf{R}\{\alpha(A_i)\}$, where we allow α to be treatment specific. This differs from the usual presentation of GEE, in which $\mathbf{V}_i$ is constant and does not depend on A_i. Because our model does not place any restrictions on $\mathbf{V}_i$, we generalize the usual approach to allow $\mathbf{V}_i$ to be more flexible. Variance parameters ϕ and α_k, where k indexes treatment, are estimated by the method of moments using $\hat{\beta}_{init}$, an initial estimator of β. To recover the GEE fit in standard software,

the above expressions simplify such that $v(A_i) = v(1) = v(0) = \mathbf{V}$, and a single correlation parameter α is estimated across all clusters. In a slight abuse of notation, we take $\mathbf{V}_i$ to be the matrix function $v(A_i)$ and $\mathbf{V}$ the constant variance matrix.

For continuous outcomes with the identity link $\mathbf{g}(A_i;\beta) = \mathbf{A}_i^*\beta$, where $\mathbf{A}_i^*$ is the $n_i \times 2$ design matrix composed of rows (1, A_i), the solution to 1.2, $\hat{\beta}$, exists in closed form, with

$$\hat{\beta} = \left(\sum_{i=1}^{m} \mathbf{A}_i^{*\mathrm{T}} \mathbf{V}_i^{-1} \mathbf{A}_i^*\right)^{-1} \left(\sum_{i=1}^{m} \mathbf{A}_i^{*\mathrm{T}} \mathbf{V}_i^{-1} \mathbf{Y}_i\right). \tag{1.3}$$

For simple designs, a closed form solution for $\hat{\beta}$ can also be derived for non-identity link functions $g(A_i;\beta)$ using the discreteness of A. The solution to 1.2 for the logit link is given in Appendix A. Generally, for more complex models, GEE coefficient estimates are found using an iterative procedure such as the Newton-Raphson method or Iteratively Reweighted Least Squares (IRWLS).

Robins et al. (1994) and Robins (1999) established that in a model for data $O = (Y, A, X)$ in which $\pi_k = P(A = k|X)$ is known, any regular and asymptotically linear estimator for β can be found as the solution to $\sum_i \psi_{i_{aug}}(Y, A, X; \bar{\gamma}_K) = \mathbf{0}$ for a specific choice of $\bar{\gamma}_K(X)$. Zhang et al. (2008) demonstrated use of this theory in RCTs with univariate outcomes. Applying these results to multivariate settings, $\hat{\beta}$ may be improved by augmenting the Standard GEE with a function of baseline covariates X. The general form of the augmented GEE for a K-level treatment is

$$\sum_{i=1}^{m} \psi_{i_{aug}}(\mathbf{Y}, A, \mathbf{X}; \beta, \bar{\gamma}_K) = \sum_{i=1}^{m} \left[\mathbf{D}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \big\{\mathbf{Y}_i - \mathbf{g}(A_i;\beta)\big\} - \sum_{k=1}^{K} \{A_{k_i} - \pi_k\} \gamma_k(\mathbf{X}_i)\right] = \mathbf{0}, \tag{1.4}$$

where $A_{k_i} = I(A_i = k)$and $\gamma_k(\mathbf{X}_i)$ is a p-dimensional function of $\mathbf{X}_i$.

It was further shown that for the class of estimating functions $\{\psi_{aug}(\bar{\gamma}_K) : \bar{\gamma}_K \in \Gamma_K\}$, where Γ_K is the set of all functions of X such that $E[\psi_{aug}(\bar{\gamma}_K)^{\mathrm{T}} \psi_{aug}(\bar{\gamma}_K)] < \infty$, the optimal estimator within this class for a fixed $\psi(Y, A; \beta)$ is obtained by setting $\gamma_{k_{opt}}(X_i) = E\{\psi_i(Y, A; \beta)|A_i = k, X_i\}$\{Robins et al. (1994); Robins (1999); Zhang et al. (2008)\}. When only two treatment arms are considered, the augmentation term

$\sum_{k=1}^{K}\{I(A_i=k)-\pi_k\}\gamma_k(\mathbf{X}_i)$ can be written as $(A_i-\pi_1)\big[\mathbf{D}_i(1)^{\mathrm{T}}\mathbf{V}_i(1)^{-1}\big\{E(\mathbf{Y}_i|A_i=1,\mathbf{X}_i)-\mathbf{g}(1;\beta)\big\}-\mathbf{D}_i(0)^{\mathrm{T}}\mathbf{V}_i(0)^{-1}\big\{E(\mathbf{Y}_i|A_i=0,\mathbf{X}_i)-\mathbf{g}(0;\beta)\big\}\big]$. The Simple Augmented GEE is thus

$$\sum_{i=1}^{m}\psi_{i_{opt}}(\mathbf{Y},A,\mathbf{X};\beta)=\sum_{i=1}^{m}\mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\big\{\mathbf{Y}_i-\mathbf{g}(A_i;\beta)\big\}-(A_i-\pi)\gamma(\mathbf{X}_i)=\mathbf{0}, \qquad (1.5)$$

where $\gamma(\mathbf{X}_i)=\big[\mathbf{D}_i(1)^{\mathrm{T}}\mathbf{V}_i(1)^{-1}\big\{E(\mathbf{Y}_i|A_i=1,\mathbf{X_i})-\mathbf{g}(1;\beta)\big\}-\mathbf{D}_i(0)^{\mathrm{T}}\mathbf{V}_i(0)^{-1}\big\{E(\mathbf{Y}_i|A_i=0,\mathbf{X}_i)-\mathbf{g}(0;\beta)\big\}\big]$.

Solving for $\hat{\beta}_{aug}$ therefore requires knowledge of ϕ, α_k, π, and $E(\mathbf{Y}_i|\mathbf{X}_i,A_i=k)$ for $k=0,1$. Following standard practice, we estimate ϕ and α_k using the residuals from a GLM fit under independence. Specifically, ϕ is estimated by the Pearson Chi-Square statistic, and α_k is obtained by solving the treatment-specific moment equations $\sum_{i=1}^{m}I(A_i=k)\big\{\hat{\epsilon}_{ij}\hat{\epsilon}_{ij}-h(\alpha_k)\big\}=0$, where $\hat{\epsilon}_{ij}=Y_{ij}-g(A_i;\hat{\beta}_{init})$, and $h(\alpha_k)$ is determined by the correlation structure assumed.

For fixed $\psi_i(\mathbf{Y},A;\beta)$, the optimality of the augmentation depends on correct estimation of $E(\mathbf{Y}_i|\mathbf{X}_i,A_i=k)$=$f_k(\mathbf{X}_i;\eta_k)$. When $E(\mathbf{Y}_i|\mathbf{X}_i,A_i=k)$ is misspecified, asymptotic normality and consistency hold, but the resulting estimator does not achieve maximum asymptotic efficiency. Several options are available for estimating the conditional mean $E(\mathbf{Y}_i|\mathbf{X}_i,A_i)$. We propose a strategy below which in large samples is guaranteed to improve on Standard GEE. Following Tsiatis' approach of estimating $E(\mathbf{Y}_i|\mathbf{X}_i,A_i)$ separately within each arm, estimation proceeds via ordinary least squares (OLS), or maximum likelihood (ML) on an appropriately defined generalized linear model. Although the observations within a cluster are not independent, the predicted values from OLS and ML fits remain consistent. For treatment-specific estimation, the argument in Leon et al. (2003) may be generalized to GEE, guaranteeing that when $E(\mathbf{Y}_i|\mathbf{X}_i,A_i)$ is estimated with OLS, the augmented estimator is at least as efficient as the unaugmented estimator for continuous and discrete outcomes. This property holds even if models are misspecified. To more correctly specify the mean function, one may opt to fit an appropriate GLM, such as logistic regression for a binary outcome. We explore both approaches through

simulation. It is also worthwhile to note that if the probability of treatment depends on baseline covariates $\mathbf{X}_i$ such that $\pi_k = P(A_i = k|\mathbf{X}_i)$, the Simple Augmented GEE does provide a valid estimate of treatment effects, but OLS is no longer sufficient to guarantee efficiency improvement over unaugmented methods. For continuous Y_{ij} and identity link $g(A_i;\beta)$, the improved estimator is

$$\hat{\beta}_{aug} = \left[\sum_{i=1}^{m} \mathbf{A}_i^{*\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{A}_i^* - (A_i-\pi)\big\{\mathbf{A}_i^*(1)^{\mathrm{T}}\mathbf{V}_i(1)^{-1}\mathbf{A}^*(1) - \mathbf{A}_i^*(0)^{\mathrm{T}}\mathbf{V}_i(0)^{-1}\mathbf{A}^*(0)\big\}\right]^{-1} \times \left[\sum_{i=1}^{m} \mathbf{A}_i^{*\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{Y}_i - (A_i-\pi)\big\{\mathbf{A}_i^*(1)^{\mathrm{T}}\mathbf{V}_i(1)^{-1}\hat{E}(\mathbf{Y}_i|A_i=1,\mathbf{X}_i) - \mathbf{A}_i^*(0)^{\mathrm{T}}\mathbf{V}_i(0)^{-1}\hat{E}(\mathbf{Y}_i|A_i=0,\mathbf{X}_i)\big\}\right]. \quad (1.6)$$

As in the unaugmented case, a closed form solution can be derived for non-identity links under a simple design. Solutions for the logit link may be found in Appendix A.

Implementation of the Simple Augmented GEE for inclusion of baselines covariates in analysis of a cluster randomized trial is summarized in the following steps:

1. Determine $\hat{E}(\mathbf{Y}_i|\mathbf{X}_i, A_i = k) = f_k(\mathbf{X}_i;\hat{\eta}_k)$ from OLS or ML regression of Y onto baseline covariates X within each treatment arm.
2. Fit a GLM under independence to obtain $\hat{\beta}_{init}$.
3. Estimate ϕ and α_k from $\hat{\epsilon}_{ij}$ of the initial fit.
4. Construct the augmented estimating equations $\psi_{aug}(\mathbf{Y}, A, \mathbf{X};\beta)$.
5. Solve for $\hat{\beta}_{aug}$.

The GEE was initially proposed as an iterative procedure, in which fitting involved repeatedly estimating correlation parameters α and mean parameters β until convergence. Since its inception, however, theoretical development and simulation studies have shown that the one-step procedure, as we have proposed for the augmented estimator, provides asymptotically equivalent estimates to the fully iterated approach, with similar finite sample properties {Lipsitz et al. (1994)}.

1.3 Variance Estimation

The asymptotic variance of $\hat{\beta}_{aug}$, under $m \to \infty$, is derived through the usual M-estimator Taylor expansion, accounting for the nuisance parameters $\hat{\eta}_k$ involved in estimating $E(Y|X, A = k)$=$f_k(\mathbf{X}_i; \eta_k)$. We let $\hat{\psi}_{i_{opt}}(\mathbf{Y}, A, \mathbf{X}; \beta)$ be an estimate of (1.5) evaluated at $\hat{\eta}$. The familiar sandwich variance estimator $var(\hat{\beta}_{aug}) = \Gamma^{-1}\Delta\Gamma^{-1^{\mathrm{T}}}$ is obtained, where $\Gamma = E[\frac{\partial\psi_{i_{opt}}(\mathbf{Y},A,\mathbf{X};\beta)}{\partial\beta^{\mathrm{T}}}]$, and $\Delta = E[\psi_{i_{opt}}(\mathbf{Y}, A, \mathbf{X}; \beta)^{\otimes 2}]$, where $U^{\otimes 2} = UU^{\mathrm{T}}$. By randomization, the augmentation term has mean zero and does not contribute to Γ. We therefore estimate Γ by $\hat{\Gamma} = m^{-1}\sum_i \mathbf{D_i}^{\mathrm{T}}\mathbf{V_i}^{-1}\mathbf{D_i}$. Estimation of η_k results in additional terms in our expansion of $\hat{\psi}_{i_{opt}}(\mathbf{Y}, A, \mathbf{X}; \beta)$ shown below.

$$\sum_{i=1}^{m} \hat{\psi}_{i_{opt}}(\mathbf{Y}, A, \mathbf{X}; \beta) =$$
$$\sum_{i=1}^{m} \Bigg\{ \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\big\{\mathbf{Y}_i - \mathbf{g}(A_i;\beta_0)\big\} - (A_i - \pi)\times$$
$$\Big[\mathbf{D}_i(1)^{\mathrm{T}}\mathbf{V}_i(1)^{-1}\big\{f_1(\mathbf{X}_i;\hat{\eta}_1) - \mathbf{g}(1;\beta_0)\big\} - \mathbf{D}_i(0)^{\mathrm{T}}\mathbf{V}_i(0)^{-1}\big\{f_0(\mathbf{X}_i;\hat{\eta}_0) - \mathbf{g}(0;\beta_0)\big\}\Big]\Bigg\} \tag{1.7}$$

$$= \sum_{i=1}^{m} \tilde{\psi}_{i_{opt}} = \sum_{i=1}^{m} \Bigg\{ \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\big\{\mathbf{Y}_i - \mathbf{g}(A_i;\beta_0)\big\}$$
$$- (A_i - \pi)\Big[\mathbf{D}_i(1)^{\mathrm{T}}\mathbf{V}_i(1)^{-1}\big\{f_1(\mathbf{X}_i;\eta_1^*) - \mathbf{g}(1;\beta_0)\big\} - \mathbf{D}_i(0)^{\mathrm{T}}\mathbf{V}_i(0)^{-1}\big\{f_0(\mathbf{X}_i;\eta_0^*) - \mathbf{g}(0;\beta_0)\big\}\Big] \tag{a}$$
$$- (A_i - \pi)\Big[\mathbf{D}_i(1)^{\mathrm{T}}\mathbf{V}_i(1)^{-1}\big\{f_1'(\mathbf{X}_i;\eta_1^*)\big\}(\hat{\eta}_1 - \eta_1^*) - \mathbf{D}_i(0)^{\mathrm{T}}\mathbf{V}_i(0)^{-1}\big\{f_0'(\mathbf{X}_i;\eta_0^*)\big\}(\hat{\eta}_0 - \eta_0^*)\Big]\Bigg\} \tag{b}$$
$$+ o_p(1)$$

where $\hat{\eta}_k \xrightarrow{P} \eta_k^*$. By randomization (b) $\xrightarrow{P} 0$ as $m \to \infty$, showing that asymptotically there is no additional variability associated with estimating η_k, even when $f_k(\mathbf{X}_i; \eta_k)$ is misspecified. For cluster randomized designs, however, the asymptotics may not hold, as the number of experimental units may be fairly small. In small sample settings, it is likely that $var(\hat{\beta}_{aug})$ is affected by estimation of η_k. We therefore estimate Δ by

$\hat{\Delta} = m^{-1}\sum_i \tilde{\psi}_{i_{opt}}^{\otimes 2}$, with and without term (b) and evaluate our variance estimator through simulation. Specifically, we estimate $(\hat{\eta}_k - \eta_k^*)$ by its first order approximation and substitute estimated parameter values for the truth. Inclusion of (b) is not guaranteed to increase the estimated variance but does provide a more unbiased estimate by accounting for estimation of η_k.

The sandwich variance estimator of Standard GEE is known to often be biased downward for inference involving relatively few independent units. We examine Fay's bias-correction approach to recover loss. We choose this approach because unlike other methods which were derived for Standard GEE, Fay's method is generalizable to any M-estimator, including our augmented estimating equations. We apply Fay's first correction, in which Δ is estimated by $\hat{\Delta}^* = m^{-1}\sum_i (\mathbf{H}_i\hat{\psi}_i)^{\otimes 2}$, where $\mathbf{H}_i$ is a diagonal matrix with $H_{i_{jj}} = \left[1 - min\{q, (\frac{\partial\psi_i(\mathbf{Y},A,\mathbf{X};\beta)}{\partial\beta^{\mathrm{T}}} \times \hat{\Gamma})_{jj}\}\right]^{-1/2}$. Lower bound q is typically set to 0.75 to prevent gross inflation {Fay and Graubard (2001)}. In total, we consider four standard error estimators for the Simple Augmented GEE: 1) unadjusted sandwich (SE_1), 2) nuisance ($\hat{\eta}_k$)-adjusted sandwich [term (b)] (SE_2), 3) sandwich with Fay's small-sample bias correction (SE_3), and 4) sandwich with Fay's small-sample bias correction and nuisance adjustment (SE_4), and evaluate each through simulation. Formulas for each estimator are provided in Appendix B.

An alternative estimate $\hat{var}(\hat{\beta}_{aug})$ can be computed through a resampling technique such as the nonparametric bootstrap. To preserve the number of treated and control clusters within any bootstrap sample, we resample clusters within treatment arm. We use strategy 1 described by Davidson and Hinkley (1997), in which the composition of resampled clusters is maintained, and demonstrate this approach through data analysis.

1.4 Application: *YOUNG CITIZENS* Study

We applied the Simple Augmented GEE to data from the *Young Citizens* study Kamo et al. (2008). This trial involved a behavioral intervention designed to train chil-

dren ages 10-14 to educate their communities about HIV. To facilitate randomization, 30 communities, were grouped into 15 pairs using a clustering algorithm involving several demographic characteristics. One community per pair was randomly assigned to treatment and the other to control. Residents within each community were surveyed post-intervention regarding their beliefs about the ability of children to effectively teach their peers and families about HIV. The primary outcome was a composite score reflecting the strength of this belief (Y_1). A secondary outcome measured residents' beliefs regarding whether or not the AIDS problem was getting worse in their communities. Residents responded on a 4-point scale with values 'strongly disagree', 'disagree', 'agree', and 'strongly agree'. Responses were dichotomized by collapsing 'strongly agree' and 'agree' into one category labeled 'agree'; 'strongly disagree' and 'disagree' were collapsed similarly (Y_2). The number of residents surveyed per community ranged from 16-80 by multiples of 16.

We implemented Standard and Augmented GEE using the customary single correlation parameter as well as the less restrictive approach of allowing treatment-specific correlation. For augmented estimators, we also included estimation under independence to further examine the impact of covariance selection on the efficiency of augmented inference. Adjustment models were determined separately for treatment and control groups by various model selection procedures. The final models used in analysis were selected via cross validation. For child efficacy (Y_1), the adjustment model in the treatment arm included the baseline covariates employment status, residential or urban community, the number of relatives living in the community, age, religion, population density, and whether or not the household had a flushing toilet, which was an indicator of household wealth. Among control communities, employment, age, and flushing toilet were included. The baseline covariates that entered the adjustment models for beliefs about the state of the HIV problem (Y_2) were mean community wealth, ethnic group, and household wealth for the intervention arm, and only mean community wealth for the control arm.

In analyzing our continuous outcome Y_1, we evaluated the marginal treatment effect by considering model (1.1), where $g(A_i;\beta)$ was the identity link function. We computed the standard error of $\hat{\beta}_1$ by the sandwich estimator and the nonparametric bootstrap for each estimation procedure. The standard error modifications in section 3 were applied, namely: 1) unadjusted sandwich (SE_1), 2) nuisance ($\hat{\eta}$)-adjusted sandwich (term (b) above) (SE_2), 3) sandwich with small-sample bias correction (SE_3), and 4) sandwich with bias correction and nuisance adjustment (SE_4). In our second application, we evaluated the marginal treatment effect on the binary secondary outcome Y_2 and fit model (1.1) with the inverse logit link. We compared estimates obtained from Standard GEE, the Simple Augmented GEE, adjusted logistic GEE with standardization i.e. the G-formula Robins (1986), and inverse probability of treatment weighted (IPTW) methods. In the IPTW approach, we ignore that the treatment probability is known and estimate $P(A=1|X)$ using a logistic regression model in which covariates were entered linearly.

In standard and augmented analyses, the intervention had a highly significant impact on the perceived ability of children to be peer educators {95% CI Standard (0.182, 0.526), 95% CI Augmented (0.245, 0.482)}. The adjusted sandwich variance estimator suggested over a 70% increase in efficiency resulting from covariate adjustment {RE, Table 1.1(a)}. Bootstrap estimates showed a similar efficiency gain under common correlation (58%) and a much more modest gain using treatment specific correlation (5%) {'RE boot', Table 1.1(b)}. Comparing within unaugmented and augmented estimators, little difference in standard error was observed between estimators allowing for treatment-specific correlation versus estimators assuming common correlation. Estimates of β_1 were similar across standard and augmented estimators with either correlation structure.

Examining our binary outcome, Y_2, there was a marked difference in the estimated parameters when comparing standard and augmented GEE (Table 1.2). The estimate $\hat{\beta}_1$ was -0.238 {95% CI (-0.777, 0.300) } using standard methods, compared to values in the range (-0.079,-0.023) for all augmented GEE estimates. In either approach, the effect of treatment on the perception of the AIDS epidemic was not significant at the p=0.05 level

Table 1.1: **Marginal Treatment Effect Analysis: Parameter estimates, sandwich standard errors, and bootstrap standard error**. *Std: unaugmented GEE, Aug: augmented GEE. Ind: independence, Exch: exchangeable correlation with single parameter, Exch-TS: exchangeable with treatment-specific correlation parameters. SE_1: unadjusted sandwich SE, SE_2: sandwich with nuisance parameter adjustment, SE_3: sandwich with small-sample adjustment, SE_4: sandwich with small-sample and nuisance adjustments. Exch: Exchangeable correlation with single parameter, Exch-TS: Exchangeable with treatment-specific correlation parameters. RE: Relative efficiency, square of the sandwich SE of the Std(Exch) estimator divided by the square of the sandwich SE of the indicated estimator. RE boot: bootstrap relative efficiency, square of the bootstrap SE of the Std(Exch) estimator divided by the square of the bootstrap SE of the indicated estimator. RE and Confidence intervals(CIs) are based on SE_3 and SE_4 for unaugmented and augmented estimators, respectively.*

(a)

	Estimate	Sandwich Standard Error with Adjustments					
Estimator	$\hat{\beta}_1$	SE_1	SE_2	SE_3	SE_4	95% CI	RE
Std (Exch)	0.354	0.082	-	0.088	-	(0.182,0.526)	1.000
Std (Exch-TS)	0.355	0.082	-	0.086	-	(0.186,0.525)	1.034
Aug (Ind)	0.360	0.066	0.060	0.071	0.064	(0.236,0.485)	1.528
Aug (Exch)	0.364	0.063	0.057	0.067	0.060	(0.245,0.482)	1.730
Aug (Exch-TS)	0.360	0.063	0.061	0.066	0.064	(0.234,0.485)	1.769

(b)

	Bootstrap SE	
Estimator	SE boot	RE boot
Std (Exch)	0.081	1.000
Std (Exch-TS)	0.085	0.910
Aug (Ind)	0.062	1.681
Aug (Exch)	0.064	1.582
Aug (Exch-TS)	0.082	0.966

{95% CI Augmented GEE=(-0.491,0.332)}. Estimates from the standardized adjusted logistic GEE were also closer to 0. Although effects were not significant at the p=0.05 level for any of the approaches, confidence intervals for the augmented GEE were somewhat tighter, as were those using standard methods of covariate adjustment.

Considering efficiency, for both outcomes the estimated variability was lower for the augmented estimator compared to Standard GEE. Although there is some uncertainty regarding the behavior of the sandwich estimator in small samples, these results suggest that when the asymptotics hold, augmented GEE is a valid approach that may be substan-

Table 1.2: **Marginal Treatment Effect Analysis with Binary Outcome: Parameter sandwich standard error estimates.** *Std: unaugmented GEE, Aug: augmented GEE. Ind: independence, Exch: exchangeable correlation with single parameter, Exch-TS: exchangeable with treatment-specific correlation parameters. GLM, OLS: generalized linear model or ordinary least squares augmentation. RE: Relative efficiency, square of the sandwich SE of the Std(Exch) estimator divided by the square of the sandwich SE of the indicated estimator. Confidence intervals (CIs) and RE based on adjusted sandwich standard errors. Adjusted Logistic GEE-Model 1:* $logit(P(Y_{ij} = 1)) = \eta_0 + \eta_1 Mean\_wealth_i$, *Model 2:* $logit(P(Y_{ij} = 1)) = \eta_0 + \eta_1 Mean\_wealth_i + \eta_2 I(Ethnic_{ij} = 1) + \eta_3 I(Wealth_{ij} = 0)$. *IPTW:* $logit(P(A_i = 1)) = \eta_0 + \eta_1 Know\_leader_i + \eta_2 Good\_floor_i$.

Estimator	$\hat{\beta}_1$	SE	95%CI	RE
Standard GEE				
Std Exch	-0.238	0.275	(-0.777,0.300)	1.000
Std Exch TS	-0.219	0.266	(-0.74,0.301)	1.069
Augmented GEE				
Aug Ind-GLM	-0.062	0.215	(-0.484,0.361)	1.627
Aug Exch-GLM	-0.079	0.21	(-0.491,0.332)	1.716
Aug Exch-TS-GLM	-0.065	0.204	(-0.465,0.335)	1.811
AugInd-OLS	-0.023	0.22	(-0.454,0.408)	1.561
AugExch-OLS	-0.062	0.214	(-0.481,0.358)	1.648
AugExch-TS-OLS	-0.047	0.206	(-0.451,0.358)	1.773
Adjusted Logistic GEE				
Model 1	-0.093	0.167	(-0.420,0.234)	2.712
Model 2	-0.044	0.179	(-0.396,0.308)	2.349
IPTW Logistic GEE				
Model 3	-0.293	0.236	(-0.756,0.170)	1.354

tially more efficient than Standard GEE. Randomized trials involving longitudinal data with many subjects or clustered designs with many smaller units, such as households or offspring, are therefore ideal candidates for this method. We evaluate our method and the behavior of the sandwich variance estimator through simulation in the following section.

1.5 Simulation Study

We assessed the performance of the Simple Augmented GEE in two sets of simulations. The first investigates continuous outcomes with an identity link, and the second set explores binary outcomes using an inverse logit link. We considered the impact of

misspecification of the augmentation term and working covariance structure on the performance of our estimator. Results are based on 1000 simulated datasets.

1.5.1 Continuous Outcomes

Cluster level covariates were treatment, density, wealth, and community type (eg., urban/rural, etc). Treatment assignment was completed by first fixing the number of treated and control clusters to $m/2$, where m is the total number of clusters. Clusters were then randomly assigned to treatment or control. Community type was generated from a multinomial distribution. Density and wealth were generated from the exponential and normal distributions, respectively. Individual-level covariates age, employment, security1, and security2 were simulated from normal and multinomial distributions with age treated as continuous and other covariates categorical. Data were generated following the means and variances of covariates in the *Young Citizens* data. Intracluster correlation was induced by cluster-specific random effects and community-level covariates. We considered varying levels of correlation for treated versus control clusters. To assess small-sample performance, we compared scenarios of $m = 30$ and $m = 100$ clusters.

Outcomes were generated from the following models:

$(Y_{ij}|X_{ij}, A_i = 1) = 7.23 + 0.599employed_{ij} + 0.44mean\_wealth_i - 0.22I(security1_{ij} = 3) - 0.06age_{ij} + 49.702density_i + b_{1i} + \epsilon_{ij}$, and $(Y_{ij}|X_{ij}, A_i = 0) = 2.56 + 0.245employed_{ij} + 0.691I(community\_type_i = 4) + 0.921I(security2_{ij} = 4) + 0.055age_{ij} + b_{0i} + \epsilon_{ij}$, where $b_{ki} \sim N(0, \sigma_k^2)$, and $\epsilon_{ij} \sim N(0, \sigma^2)$. Community-level and individual-level covariates therefore contributed to heterogeneity in subject responses. Values of σ_1^2 and σ_0^2 were selected to yield the desired within-cluster marginal correlation (ρ_k). For treatment and control clusters alike, $\sigma^2 = 1$.

We evaluated the effect of working covariance and augmentation misspecification by estimating β_1 and its variance under different covariance structures and augmentation models. Two variations of Standard GEE were considered: Standard GEE with com-

mon exchangeable correlation {Std(Exch)}, and Standard GEE with treatment-specific exchangeable correlation {Std(Exch-TS)}. For the class of augmented GEE, we estimated $\hat{\beta}_{aug}$ with independence, exchangeable, and treatment-specific exchangeable correlation structures. Each estimator was evaluated under several augmentation models. The estimator resulting from fitting the true form of $E(Y_{ij}|X_{ij}, A_i = k)$ is denoted by 'C' for 'Correct'. Alternative augmentation models were defined by forward (F) and backward (B) selection, and a wrong (W) model. The wrong models were given by $E(Y_{ij}|X_{ij}, A_i = 1) = \eta_0 + \eta_1 mean\_wealth_i + \eta_2 I(community\_type_i = 2) + \eta_3 I(security1_{ij} = 2)$ and $E(Y_{ij}|X_{ij}, A_i = 0) = \eta_0 + \eta_1 density_i + \eta_2 age_{ij} + \eta_3 I(community\_type_i = 1) + \eta_4 I(security2_{ij} = 4)$. Augmentation under the 'Correct' model illustrates the largest possible efficiency gain. Alternative model fitting techniques were chosen to be representative of methods commonly used when performing covariate adjustment in analyzing clinical trial data. Forward or backward stepwise selection may be used by analysts favoring more parsimonious or larger models, respectively. The 'Wrong' model was included for comparison using models that contain some relevant covariates but omit others. To correct for small-sample variance underestimation, we applied several modifications to the sandwich estimator as detailed in Section 1.3. For the unaugmented estimators, we calculated the sandwich variance (SE_1), and the sandwich variance with bias correction (SE_3). Standard errors for augmented estimators were calculated using SE_2 and SE_4 as well, which account for η_k-estimation.

Table 1.3 shows results for $m = 30$ and 100, $\sigma_1^2 = 0.03$, and $\sigma_0^2 = 0.025$, which correspond to approximately 10% and 5% within-cluster correlation in treated and control clusters, respectively. For Table 1.4, we raise the level of unexplained similarity among cluster members by setting $\sigma_1^2 = 0.23$ and $\sigma_0^2 = 0.20$ for the sample sizes previously considered.

For small-sample and large-sample inference, bias was similar across all estimators. Working covariance specification affected the variance of the augmented estimator, with exchangeable (true) correlation structures resulting in smaller average standard errors than independence. Comparing estimators calculated with an exchangeable corre-

lation structure, augmented estimators were often more efficient than the standard approach. Monte Carlo relative efficiency estimates suggest that in the small-sample setting with low levels of unexplained intracommunity correlation, considerable improvement (5-19%) is observed even when misspecifying the augmentation model (Table 1.3). When intracluster correlation was larger, additional variability associated with automated model selection resulted in loss of efficiency associated with augmentation (Table 1.4). Average sandwich standard errors were overly optimistic in comparing augmented GEE to Standard GEE in small samples, consistently estimating lower variability with augmentation. For large samples, efficiency gains were not hindered by higher levels of unexplained cluster similarity, with Monte Carlo efficiency improving by 5-40% (Table 1.4).

Table 1.3: **Standard vs. Augmented GEE, Continuous Outcome: 30 & 100 clusters,** $\rho_0 = 0.05$**,** $\rho_1 = 0.10$**,** $\beta_1 = 1.3239$**.** *Std: unaugmented. C,F,B,W: augmented with 'Correct', 'Forward' selected, 'Backward' selected, or 'Wrong' models. Ind: independence, Exch: exchangeable with single correlation parameter. Exch-TS: exchangeable with treatment-specific parameter.* SE_1*: average unadjusted sandwich SE,* SE_2*: average sandwich SE with nuisance parameter adjustment,* SE_3*: average sandwich SE with small-sample adjustment,* SE_4*: average sandwich SE with small-sample and nuisance adjustments. MC SE: Monte Carlo standard deviation. MC RE: square of the Monte Carlo SE of the Std(Exch) estimator divided by the Monte Carlo SE for the indicated estimator. Cov. U:* SE_1 *coverage, Cov. A:* SE_3 *and* SE_4 *coverage for unaugmented and augmented GEE, respectively.*

m=30	Bias	SE_1	SE_2	SE_3	SE_4	MC SE	MC RE	Cov. U	Cov. A
Std(Exch)	0.000	0.156	-	0.165	-	0.165	1.000	0.933	0.949
Std(Exch-TS)	-0.001	0.158	-	0.166	-	0.164	1.004	0.934	0.949
C(Ind)	-0.001	0.135	0.138	0.140	0.144	0.151	1.194	0.912	0.929
F(Ind)	-0.001	0.132	0.136	0.137	0.142	0.153	1.162	0.895	0.922
B(Ind)	-0.001	0.132	0.137	0.137	0.142	0.154	1.149	0.893	0.918
W(Ind)	0.003	0.143	0.151	0.150	0.158	0.160	1.057	0.911	0.937
C(Exch)	-0.004	0.130	0.132	0.134	0.137	0.147	1.248	0.909	0.927
F(Exch)	-0.004	0.128	0.131	0.132	0.135	0.149	1.219	0.898	0.917
B(Exch)	-0.004	0.128	0.131	0.132	0.135	0.150	1.205	0.893	0.915
W(Exch)	-0.002	0.139	0.145	0.144	0.150	0.156	1.110	0.904	0.928
C(Exch-TS)	-0.005	0.134	0.138	0.137	0.143	0.149	1.223	0.914	0.925
F(Exch-TS)	-0.005	0.133	0.137	0.135	0.142	0.151	1.198	0.895	0.915
B(Exch-TS)	-0.005	0.132	0.137	0.135	0.142	0.151	1.186	0.890	0.910
W(Exch-TS)	-0.004	0.142	0.148	0.146	0.154	0.159	1.078	0.900	0.927
m=100	Bias	SE_1	SE_2	SE_3	SE_4	MC SE	MC RE	Cov. U	Cov. A
Std(Exch)	-0.002	0.088	-	0.090	-	0.090	1.000	0.943	0.946
Std(Exch-TS)	-0.003	0.088	-	0.089	-	0.090	1.001	0.942	0.945
C(Ind)	-0.004	0.077	0.077	0.078	0.078	0.078	1.333	0.942	0.947
F(Ind)	-0.004	0.076	0.077	0.077	0.078	0.079	1.296	0.937	0.943
B(Ind)	-0.004	0.076	0.077	0.077	0.078	0.079	1.297	0.937	0.944
W(Ind)	-0.003	0.083	0.084	0.084	0.085	0.088	1.048	0.935	0.945
C(Exch)	-0.004	0.074	0.074	0.075	0.075	0.075	1.449	0.944	0.945
F(Exch) 6	-0.005	0.073	0.074	0.074	0.075	0.076	1.409	0.933	0.941
B(Exch)	-0.005	0.073	0.074	0.074	0.075	0.076	1.410	0.932	0.94
W(Exch)	-0.003	0.080	0.081	0.080	0.082	0.083	1.166	0.936	0.943
C(Exch-TS)	-0.004	0.074	0.074	0.074	0.075	0.074	1.452	0.946	0.95
F(Exch-TS)	-0.005	0.073	0.074	0.074	0.075	0.075	1.412	0.937	0.943
B(Exch-TS)	-0.005	0.073	0.074	0.074	0.075	0.075	1.413	0.936	0.943
W(Exch-TS)	-0.003	0.079	0.080	0.080	0.081	0.083	1.168	0.934	0.943

Table 1.4: **Standard vs. Augmented GEE, Continuous Outcome: 30 & 100 clusters,** $\rho_0 = 0.13$**,** $\rho_1 = 0.17$**,** $\beta_1 = 1.3239$**..** *Std: unaugmented. C,F,B,W: augmented with 'Correct', 'Forward' selected, 'Backward' selected, or 'Wrong' models. Ind: independence, Exch: exchangeable with single correlation parameter. Exch-TS: exchangeable with treatment-specific parameter.* SE_1*: average unadjusted sandwich SE,* SE_2*: average sandwich SE with nuisance parameter adjustment,* SE_3*: average sandwich SE with small-sample adjustment,* SE_4*: average sandwich SE with small-sample and nuisance adjustments. MC SE: Monte Carlo standard deviation. MC RE: square of the Monte Carlo SE of the Std(Exch) estimator divided by the Monte Carlo SE for the indicated estimator. Cov. U:* SE_1 *coverage, Cov. A:* SE_3 *and* SE_4 *coverage for unaugmented and augmented GEE, respectively.*

m=30	Bias	SE_1	SE_2	SE_3	SE_4	MC SE	MC RE	Cov. U	Cov. A
Std(Exch)	0.012	0.217	-	0.229	-	0.233	1.000	0.933	0.937
Std(Exch-TS)	0.011	0.217	-	0.227	-	0.233	1.000	0.929	0.934
C(Ind)	0.007	0.209	0.219	0.221	0.232	0.234	0.989	0.907	0.934
F(Ind)	0.007	0.201	0.216	0.211	0.226	0.255	0.834	0.863	0.908
B(Ind)	0.007	0.202	0.216	0.211	0.227	0.256	0.828	0.861	0.911
W(Ind)	0.014	0.213	0.226	0.226	0.240	0.250	0.864	0.897	0.922
C(Exch)	0.005	0.199	0.206	0.207	0.214	0.215	1.170	0.912	0.934
F(Exch)	0.005	0.194	0.205	0.201	0.213	0.237	0.963	0.881	0.916
B(Exch)	0.005	0.194	0.206	0.201	0.213	0.238	0.954	0.879	0.915
W(Exch)	0.010	0.204	0.214	0.213	0.224	0.232	1.008	0.910	0.94
C(Exch-TS)	0.004	0.198	0.205	0.206	0.213	0.215	1.172	0.911	0.935
F(Exch-TS)	0.004	0.193	0.205	0.200	0.212	0.237	0.964	0.884	0.914
B(Exch-TS)	0.004	0.193	0.205	0.200	0.212	0.238	0.955	0.882	0.914
W(Exch-TS)	0.009	0.203	0.214	0.212	0.223	0.232	1.008	0.909	0.938
m=100	Bias	SE_1	SE_2	SE_3	SE_4	MC SE	MC RE	Cov. U	Cov. A
Std(Exch)	0.006	0.123	-	0.125	-	0.124	1.000	0.947	0.948
Std(Exch-TS)	0.006	0.123	-	0.125	-	0.123	1.004	0.947	0.949
C(Ind)	0.002	0.121	0.123	0.123	0.125	0.124	0.997	0.943	0.949
F(Ind)	0.003	0.118	0.121	0.120	0.123	0.125	0.979	0.935	0.941
B(Ind)	0.003	0.118	0.121	0.120	0.123	0.125	0.978	0.935	0.941
W(Ind)	0.004	0.124	0.127	0.127	0.129	0.128	0.931	0.935	0.944
C(Exch)	0.005	0.113	0.114	0.114	0.116	0.116	1.128	0.939	0.943
F(Exch)	0.006	0.112	0.114	0.113	0.115	0.118	1.102	0.929	0.937
B(Exch)	0.006	0.112	0.114	0.113	0.115	0.118	1.101	0.929	0.937
W(Exch)	0.007	0.117	0.118	0.118	0.120	0.120	1.056	0.934	0.944
C(Exch-TS)	0.005	0.113	0.114	0.114	0.115	0.116	1.133	0.941	0.945
F(Exch-TS)	0.006	0.111	0.114	0.113	0.115	0.118	1.106	0.931	0.939
B(Exch-TS)	0.006	0.111	0.114	0.113	0.115	0.118	1.105	0.931	0.939
W(Exch-TS)	0.007	0.116	0.118	0.118	0.120	0.120	1.061	0.935	0.942

Coverage results show that for small samples, the uncorrected sandwich variance underestimates the variability of the augmented estimator (Tables 1.3 & 1.4). Bias correction fully recovered small-sample loss of variance for Standard GEE. For augmented estimators, correction was less effective. Coverage was slightly increased by accounting for augmentation in the sandwich variance but did not quite reach nominal levels. For large-sample inference, neither adjustment substantially increased coverage, which was already close to nominal levels for the uncorrected sandwich variance without the nuisance term.

1.5.2 Binary Outcomes

To explore the performance of the augmented GEE for clustered binary outcomes, we again generated datasets of m clusters with probability of treatment $P(A = 1) = 1/2$. Cluster-level variables X_1 and X_2 were simulated from exponential and multinomial distributions with mean 0.002 and probabilities p=(0.46, 0.27, 0.07, 0.17, 0.03), respectively. Individual-level covariates X_3, X_4, X_5, and X_6 were generated such that $(X_3, X_4) \sim Normal\left(\begin{pmatrix} 0 & 0 \end{pmatrix}, \begin{pmatrix} 4 & 6 \\ 6 & 25 \end{pmatrix}\right)$, $X_5 \sim Bernoulli(p = 0.28)$, and $X_6 \sim Multinomial\{1, p = (0.45, 0.15, 0.30, 0.10)\}$. We used the random intercept logistic model to simulate correlated binary outcomes $\mathbf{Y}$. Random intercepts b_i were drawn from the bridge distribution for the logit link {Wang and Louis (2003)}, $B_l(0, 1-\rho)$, where 0 is the mean and ρ is the desired correlation. The bridge distribution was selected to preserve the logistic shape after marginalizing over random effects and provide a simple scaling relationship between parameters of the models for $E(\mathbf{Y}|\mathbf{X}, A, b)$ and $E(\mathbf{Y}|\mathbf{X}, A)$. Outcome generating models were: $logit\{E(Y_{ij}|X_{ij}, A_i = 1, b)\} = \eta_{1_0} + \eta_{1_1} X_{3_{ij}} + \eta_{1_2} X_{4_{ij}} + b_i$, and $logit\{E(Y_{ij}|X_{ij}, A_i = 0, b)\} = \eta_{0_0} + \eta_{0_1} X_{4_{ij}} + \eta_{0_2} X_{4_{ij}}^2 + \eta_{0_3} X_{5_{ij}} + b_i$.

For low association between Y and X, we set $\eta_0 = (3.4, -0.6, 0.03, 0.5)^{\mathrm{T}}$ and $\eta_1 = (2.5, -0.62, 0.86)^{\mathrm{T}}$. Coefficients $\eta_0 = c(2.0, -0.9, 0.03, 0.5)^{\mathrm{T}}$ and $\eta_1 = (1.5, -0.62, 0.86)^{\mathrm{T}}$ were used for a high association. We again compared small-sample versus large-sample performance by implementing standard and augmented GEE under $m = 30$, $m = 100$,

and $m = 250$ clusters. We also considered two levels of intracluster correlation (ρ=0.05, 0.20). Results for $m = 250$ are included in Appendix C.

We applied the augmented GEE under independent and exchangeable correlation structures and evaluated different methods of fitting the augmentation term. To guarantee improved efficiency relative to Standard GEE, we fit augmentation models $E(\mathbf{Y}|\mathbf{X}, A = k)$ using OLS. We contrast this approach with logistic regression, which correctly specifies the form of the relationship between Y and X, but is not guaranteed to improve efficiency under model misspecification. For each model fitting technique, we fit the correct augmentation model (C), a forward selection model (F), and two wrong models (O & W). Wrong models denoted by 'O' contained one baseline covariate. Specifically, the models fit were $E(Y_{ij}|X_{ij}, A = 1) = g(\alpha_{1_0} + \alpha_{1_1}X_{4_{ij}})$ and $E(Y_{ij}|X_{ij}, A = 0) = g(\alpha_{0_0} + \alpha_{0_1}X_{3_{ij}})$. Wrong models 'W' are given by $E(Y_{ij}|X_{ij}, A = 1)\} = g(\alpha_{1_0} + \alpha_{1_1}X_{5_{ij}} + \alpha_{1_1}X_{2_i})$ and $E(Y_{ij}|X_{ij}, A = 0) = g(\alpha_{0_0} + \alpha_{0_1}X_{4_{ij}} + \alpha_{0_2}X_{1_{ij}} + \alpha_{0_3}X_{5_{ij}})$.

Results were similar to those obtained for the continuous outcomes in the first set of simulations (Tables 1.5 - 1.8). Bias was similar across all methods of estimation for small- and large-sample inference, and correct specification of the working covariance resulted in more efficient estimation for augmented estimators. Small-sample results suggested that for low association of baseline covariates and outcome, small gains are possible for reasonably specified models (2%-10%), but for automated model selection and poorly specified models, efficiency loss occurs (-17%- -3%) because of additional variability introduced by model selection and estimation of the augmentation terms. Efficiency increased by 8%-35% when baseline covariates were more strongly related to the outcome, and unexplained intracluster correlation was low. For higher intracluster correlation, efficiency gains were lower (-5% - 12%, Table 1.7), with loss of efficiency for automated model selection. Similar to the continuous outcome, standard error adjustments were partially effective in recovering nominal coverage. When 100 clusters were sampled, augmentation increased efficiency by 1%-35% for high association or low intracluster correlation. With low association between $\mathbf{X}$ and $\mathbf{Y}$ and high intracluster correlation, augmentation

decreased efficiency for poorly specified and automated models. Considering 250 clusters, augmented estimators were more efficient than unaugmented estimators across the levels of intracluster correlation, degree of $\mathbf{X}$,$\mathbf{Y}$ association, and methods of model fitting that were considered (Appendix C).

In summary, large-sample results suggest improvement with augmentation, whereas results for small-sample estimation are less consistent. Across the number of clusters evaluated, augmentation was less beneficial as the degree of intracluster correlation increased. Regarding augmentation fit, the variability of $\hat{\beta}_1$ was similar when comparing augmented estimators resulting from predictions from OLS and ML.

Table 1.5: **Standard vs. Augmented GEE, Binary Outcome: 30 & 100 clusters, low association,** $\rho = 0.05$, $\beta_1 = -0.2959$ low association, $\beta_1 = 1.1362$ high association. *Std: unaugmented. Correlation exchangeable unless denoted by 'Ind' for independence. C,F,O,W: augmentation with 'Correct', 'Forward' selected, 'One-variable', or 'Wrong' model. ML, OLS: augmentation fit with maximum likelihood or ordinary least squares.* SE_1*: average unadjusted sandwich SE,* SE_2*: average sandwich SE with nuisance parameter adjustment,* SE_3*: average sandwich SE with small-sample adjustment,* SE_4*: average sandwich SE with small-sample and nuisance adjustments. MC RE: square of the Monte Carlo SE of the Std(Exch) estimator divided by the Monte Carlo SE for the indicated estimator. Cov. U:* SE_1 *coverage, Cov. A:* SE_3 *and* SE_4 *coverage for unaugmented and augmented GEE, respectively.*

Low	Estimator	$\hat{\beta}_1$	Bias	SE_1	SE_2	SE_3	SE_4	MC SE	MC RE	Cov. U	Cov. A
m=30	Std	-0.299	0.003	0.196	-	0.209	-	0.220	1.000	0.923	0.945
	C - ML (Ind)	-0.300	0.004	0.193	0.196	0.207	0.210	0.220	1.002	0.924	0.942
	C - ML	-0.299	0.003	0.187	0.189	0.198	0.200	0.210	1.101	0.920	0.937
	C - OLS	-0.300	0.004	0.188	0.191	0.200	0.202	0.213	1.075	0.919	0.936
	F - ML	-0.298	0.003	0.178	0.180	0.187	0.189	0.225	0.956	0.880	0.901
	F - OLS	-0.302	0.006	0.179	0.183	0.188	0.192	0.226	0.948	0.878	0.906
	O - ML	-0.299	0.003	0.190	0.192	0.202	0.204	0.215	1.051	0.918	0.935
	O - OLS	-0.299	0.003	0.190	0.192	0.202	0.205	0.215	1.054	0.914	0.937
	W - ML	-0.295	-0.001	0.191	0.195	0.202	0.206	0.224	0.971	0.904	0.929
	W - OLS	-0.298	0.002	0.191	0.195	0.202	0.207	0.224	0.970	0.902	0.933
m=100	Std	-0.293	-0.003	0.115	-	0.117	-	0.116	1.000	0.944	0.947
	C - ML (Ind)	-0.292	-0.004	0.113	0.113	0.115	0.116	0.117	0.987	0.941	0.946
	C - ML	-0.293	-0.003	0.109	0.109	0.111	0.111	0.112	1.089	0.938	0.940
	C - OLS	-0.293	-0.003	0.110	0.110	0.112	0.112	0.112	1.077	0.943	0.945
	F - ML	-0.293	-0.003	0.107	0.108	0.109	0.110	0.113	1.057	0.934	0.944
	F - OLS	-0.293	-0.003	0.108	0.109	0.109	0.110	0.114	1.052	0.937	0.943
	O - ML	-0.293	-0.003	0.111	0.111	0.113	0.113	0.113	1.070	0.944	0.950
	O - OLS	-0.293	-0.003	0.111	0.111	0.113	0.113	0.113	1.070	0.943	0.948
	W - ML	-0.293	-0.003	0.113	0.114	0.115	0.115	0.116	1.014	0.935	0.945
	W - OLS	-0.293	-0.003	0.113	0.114	0.115	0.116	0.116	1.015	0.937	0.943

Table 1.6: **Standard vs. Augmented GEE, Binary Outcome: 30 & 100 clusters, high association,** $\rho = 0.05$, $\beta_1 = -0.2959$ low association, $\beta_1 = 1.1362$ high association. *Std: unaugmented. Correlation exchangeable unless denoted by 'Ind' for independence. C,F,O,W: augmentation with 'Correct', 'Forward' selected, 'One-variable', or 'Wrong' model. ML, OLS: augmentation fit with maximum likelihood or ordinary least squares.* SE_1*: average unadjusted sandwich SE,* SE_2*: average sandwich SE with nuisance parameter adjustment,* SE_3*: average sandwich SE with small-sample adjustment,* SE_4*: average sandwich SE with small-sample and nuisance adjustments. MC RE: square of the Monte Carlo SE of the Std(Exch) estimator divided by the Monte Carlo SE for the indicated estimator. Cov. U:* SE_1 *coverage, Cov. A:* SE_3 *and* SE_4 *coverage for unaugmented and augmented GEE, respectively.*

High	Estimator	$\hat{\beta}_1$	Bias	SE_1	SE_2	SE_3	SE_4	MC SE	MC RE	Cov. U	Cov. A
m=30	Std	1.137	-0.001	0.153	-	0.163	-	0.169	1.000	0.925	0.946
	C - ML (Ind)	1.132	0.004	0.136	0.138	0.146	0.149	0.151	1.254	0.935	0.948
	C - ML	1.135	0.001	0.132	0.134	0.140	0.142	0.144	1.382	0.932	0.952
	C - OLS	1.134	0.002	0.134	0.136	0.143	0.145	0.146	1.348	0.938	0.951
	F - ML	1.135	0.002	0.125	0.128	0.132	0.135	0.155	1.195	0.901	0.921
	F - OLS	1.137	-0.001	0.127	0.131	0.134	0.139	0.156	1.172	0.896	0.924
	O - ML	1.135	0.001	0.135	0.137	0.144	0.146	0.148	1.306	0.928	0.955
	O - OLS	1.135	0.001	0.136	0.139	0.146	0.148	0.150	1.278	0.930	0.950
	W - ML	1.137	-0.001	0.141	0.145	0.150	0.154	0.161	1.101	0.924	0.947
	W - OLS	1.137	-0.001	0.141	0.146	0.150	0.155	0.163	1.085	0.921	0.945
m=100	Std	1.138	-0.002	0.089	-	0.090	-	0.090	1.000	0.946	0.949
	C - ML (Ind)	1.139	-0.003	0.079	0.079	0.080	0.081	0.083	1.162	0.934	0.935
	C - ML	1.139	-0.003	0.076	0.076	0.078	0.078	0.080	1.257	0.936	0.941
	C - OLS	1.140	-0.004	0.077	0.078	0.079	0.079	0.081	1.234	0.943	0.945
	F - ML	1.138	-0.002	0.075	0.075	0.076	0.076	0.082	1.210	0.931	0.935
	F - OLS	1.140	-0.003	0.076	0.076	0.077	0.078	0.082	1.197	0.934	0.943
	O - ML	1.139	-0.003	0.078	0.078	0.080	0.080	0.081	1.222	0.943	0.954
	O - OLS	1.140	-0.004	0.079	0.079	0.080	0.081	0.082	1.203	0.946	0.952
	W - ML	1.139	-0.003	0.083	0.083	0.084	0.085	0.086	1.098	0.948	0.954
	W - OLS	1.140	-0.003	0.083	0.084	0.084	0.085	0.086	1.095	0.945	0.952

Table 1.7: **Standard vs. Augmented GEE, Binary Outcome: 30 & 100 clusters, low,** $\rho = 0.20$ $\beta_1 = -0.2164$ low association, $\beta_1 = 1.0501$ high association. *Std: unaugmented. Correlation exchangeable unless denoted by 'Ind' for independence. C,F,O,W: augmentation with 'Correct', 'Forward' selected, 'One-variable', or 'Wrong' model. ML, OLS: augmentation fit with maximum likelihood or ordinary least squares.* SE_1*: average unadjusted sandwich SE,* SE_2*: average sandwich SE with nuisance parameter adjustment,* SE_3*: average sandwich SE with small-sample adjustment,* SE_4*: average sandwich SE with small-sample and nuisance adjustments. MC RE: square of the Monte Carlo SE of the Std(Exch) estimator divided by the Monte Carlo SE for the indicated estimator. Cov. U:* SE_1 *coverage, Cov. A:* SE_3 *and* SE_4 *coverage for unaugmented and augmented GEE, respectively.*

Low	Bias	SE_1	SE_2	SE_3	SE_4	MC SE	MC RE	Cov. U	Cov.A	
m=30	Std	0.012	0.317	-	0.335	-	0.344	1.000	0.925	0.935
	C - ML (Ind)	0.019	0.328	0.332	0.354	0.358	0.373	0.855	0.908	0.934
	C - ML	0.013	0.312	0.313	0.329	0.330	0.338	1.041	0.927	0.939
	C - OLS	0.013	0.313	0.314	0.330	0.331	0.339	1.031	0.931	0.937
	F - ML	0.001	0.294	0.296	0.308	0.310	0.365	0.891	0.892	0.908
	F - OLS	0.008	0.297	0.303	0.310	0.317	0.376	0.838	0.887	0.905
	O - ML	0.015	0.314	0.315	0.331	0.332	0.338	1.036	0.924	0.935
	O - OLS	0.016	0.314	0.315	0.331	0.332	0.339	1.034	0.926	0.935
	W - ML	0.004	0.311	0.314	0.328	0.330	0.350	0.968	0.915	0.925
	W - OLS	0.009	0.312	0.315	0.329	0.332	0.352	0.958	0.912	0.926
m=100	Std	-0.003	0.183	-	0.186	-	0.188	1.000	0.937	0.943
	C - ML (Ind)	-0.003	0.192	0.193	0.196	0.197	0.201	0.868	0.938	0.943
	C - ML	-0.004	0.180	0.180	0.182	0.182	0.185	1.031	0.939	0.941
	C - OLS	-0.004	0.180	0.180	0.183	0.183	0.185	1.023	0.930	0.935
	F - ML	-0.003	0.176	0.177	0.179	0.179	0.190	0.975	0.925	0.930
	F - OLS	-0.003	0.176	0.178	0.179	0.180	0.191	0.968	0.930	0.933
	O - ML	-0.003	0.181	0.181	0.184	0.184	0.186	1.021	0.930	0.936
	O - OLS	-0.004	0.181	0.181	0.184	0.184	0.186	1.018	0.930	0.936
	W - ML	-0.004	0.181	0.182	0.184	0.185	0.188	0.996	0.937	0.943
	W - OLS	-0.003	0.181	0.182	0.184	0.185	0.188	0.995	0.935	0.943

Table 1.8: **Standard vs. Augmented GEE, Binary Outcome: 30 & 100 clusters, high association,** $\rho = 0.20$ $\beta_1 = -0.2164$ low association, $\beta_1 = 1.0501$ high association. *Std: unaugmented. Correlation exchangeable unless denoted by 'Ind' for independence. C,F,O,W: augmentation with 'Correct', 'Forward' selected, 'One-variable', or 'Wrong' model. ML, OLS: augmentation fit with maximum likelihood or ordinary least squares.* SE_1*: average unadjusted sandwich SE,* SE_2*: average sandwich SE with nuisance parameter adjustment,* SE_3*: average sandwich SE with small-sample adjustment,* SE_4*: average sandwich SE with small-sample and nuisance adjustments. MC RE: square of the Monte Carlo SE of the Std(Exch) estimator divided by the Monte Carlo SE for the indicated estimator. Cov. U:* SE_1 *coverage, Cov. A:* SE_3 *and* SE_4 *coverage for unaugmented and augmented GEE, respectively.*

High	Estimator	Bias	SE_1	SE_2	SE_3	SE_4	MC SE	MC RE	Cov. U	Cov. A
m=30	Std	-0.017	0.244	-	0.258	-	0.251	1.000	0.934	0.948
	C - ML (Ind)	-0.015	0.244	0.247	0.264	0.267	0.258	0.941	0.923	0.949
	C - ML	-0.019	0.230	0.231	0.243	0.244	0.235	1.132	0.935	0.950
	C - OLS	-0.019	0.231	0.232	0.245	0.246	0.237	1.120	0.934	0.947
	F - ML	-0.015	0.217	0.219	0.228	0.230	0.257	0.952	0.889	0.909
	F - OLS	-0.019	0.219	0.224	0.230	0.235	0.262	0.914	0.895	0.913
	O - ML	-0.024	0.233	0.234	0.247	0.247	0.243	1.061	0.931	0.941
	O - OLS	-0.025	0.234	0.235	0.247	0.248	0.244	1.052	0.928	0.943
	W - ML	-0.016	0.234	0.237	0.248	0.250	0.248	1.021	0.931	0.942
	W - OLS	-0.017	0.235	0.237	0.248	0.250	0.248	1.018	0.932	0.944
m=100	Std	-0.012	0.138	-	0.140	-	0.137	1.000	0.946	0.949
	C - ML (Ind)	-0.009	0.140	0.140	0.143	0.144	0.143	0.928	0.941	0.950
	C - ML	-0.009	0.130	0.130	0.133	0.133	0.129	1.124	0.952	0.958
	C - OLS	-0.009	0.131	0.131	0.133	0.133	0.130	1.117	0.953	0.956
	F - ML	-0.007	0.128	0.129	0.130	0.131	0.132	1.076	0.944	0.948
	F - OLS	-0.007	0.128	0.130	0.130	0.132	0.133	1.070	0.946	0.953
	O - ML	-0.010	0.132	0.132	0.134	0.134	0.131	1.096	0.947	0.952
	O - OLS	-0.010	0.132	0.132	0.134	0.134	0.131	1.095	0.950	0.953
	W - ML	-0.011	0.134	0.135	0.136	0.137	0.135	1.041	0.943	0.949
	W - OLS	-0.011	0.134	0.135	0.136	0.137	0.134	1.044	0.945	0.951

1.6 Discussion

This paper demonstrates the use of methodology based on semiparametric theory to improve efficiency of inferences in randomized studies with correlated outcomes through augmenting the Standard GEE. This method extends the work of Zhang et al. (2008) by focusing on multivariate outcomes, and is the first application of this approach to a cluster randomized trial.

The binary outcome analysis illustrates an additional advantage of augmented GEE - double robustness. Results from Standard GEE may result in misleading estimates in settings where randomization has led to imbalance in important predictors. Augmentation involves specifying a conditional model $E(\mathbf{Y}|\mathbf{X}, A)$ that corrects for imbalances and therefore recovers unbiased estimates of treatment effects, even when randomization does not result in independence of $\mathbf{X}$ and A in the observed data. Alternative methods for correction, such as IPTW using a predictive model for the probability of treatment given baseline covariates, may not perform well given the cluster-level assignment. Predictive models for treatment only make use of cluster-level information; individual-level covariates may be averaged by cluster to create cluster-level covariates, but this data-coarsened approach can lead to poorly specified models. Generally, inference on the probability of treatment will be poor given the small number of randomized units. Alternatively, augmentation exploits relationships among individual-level covariates and outcomes. Since there are multiple individuals per cluster, there is more information available for estimating $E(\mathbf{Y}|\mathbf{X}, A)$ compared to $P(A = 1|\mathbf{X})$. Estimation of $E(\mathbf{Y}|\mathbf{X}, A)$ may consequently result in a better estimator of β_1.

Simulation studies explored the possibility of efficiency gains using the augmented GEE in small- and large-sample settings. For large samples, the augmented GEE improved efficiency compared to the Standard GEE for marginal treatment effects, which ignores baseline covariates. In the small-sample setting, efficiency gain was less consistent; low levels of between-community heterogeneity and high degrees of association between

baseline covariates and outcomes were required to benefit from augmentation. Gail et al. (1996) found a similar trend in their studies of permutation inference, noting that covariate adjustment did not improve efficiency when between-community variability was high. These results highlight the importance of measuring all covariates that contribute to within-community similarities in response. Interpreting the results from the *Young Citizens* study using the insight obtained through simulations, the low intracluster correlation (0.02) suggests improvement in efficiency when adjusting for baseline covariates. The degree of improvement, however, may be overstated by sandwich standard errors. Small-sample estimation also resulted in coverage slightly below nominal levels, even after standard error adjustment. The standard error modifications used only consider first-order approximations to the sandwich variance and nuisance parameter distributions. The simulation results suggest second order effects of nuisance parameter estimation may impact variance underestimation. The shortcomings of this approach in small samples motivate investigation into the use of augmented estimators with permutation-based inference.

We implemented augmentation using separate models for treatment and control, with ML and OLS for binary outcomes, and OLS for continuous outcomes. Asymptotically, treatment-specific OLS including an intercept term is guaranteed to be at least as efficient as the unadjusted estimator {Leon et al. (2003); Tsiatis et al. (2008)}. As discussed by Zhang and Gilbert, data splitting can be inefficient in finite samples compared to fitting a common model for $E(\mathbf{Y}|\mathbf{X}, A)$. For studies involving relatively few randomized units, fitting a common conditional model may better utilize covariate information. The effect of data splitting in finite sample inference has not yet been examined in practice. To guarantee efficiency gain over unadjusted methods when fitting a common model, van der Laan's empirical efficiency maximization approach {Rubin and van der Laan (2008)} may be used. This method estimates nuisance parameters by empirically minimizing the asymptotic variance of a scalar targeted parameter. It results in fitting adjustment models with a weighted least squares procedure, in which weights depend on treatment probabilities.

Although the Simple Augmented GEE improves estimation in large samples, it is not the semiparametric efficient estimator for our restricted mean model for multivariate outcome data, even under correct specification of $E(\mathbf{Y}|A, \mathbf{X})$. Nonetheless, the Simple Augmented GEE builds upon Standard GEE in an intuitive way and provides insight into how augmentation may be used with multivariate data to improve efficiency. Development of a locally semiparametric efficient estimator for restricted mean models for multivariate data and an understanding of its behavior remain important research questions. A locally efficient estimator is an estimator that remains consistent and asymptotically normal under the restricted mean model, and that achieves the semiparametric efficiency bound for the model at the submodel where nuisance parameters are correctly specified. When model misspecification of nuisance parameters is present, it is not clear whether the locally efficient estimator will still improve efficiency compared to standard techniques. Additional modification of the locally efficient estimator is needed to ensure improvement relative to Standard GEE. Further research is warranted in this area.

Locally Efficient Estimation of Marginal Treatment Effects Using Auxiliary Covariates in Randomized Trials with Correlated Outcomes

Alisa J. Stephens, Eric J. Tchetgen Tchetgen, and Victor De Gruttola

Department of Biostatistics

Harvard University

2.1 Introduction

Semiparametric estimators are appealing for their robustness to distributional assumptions and model misspecification. In the analysis of randomized trials, semiparametric theory has been used to develop estimators of treatment effects that improve efficiency of inferences by incorporating baseline covariates, where 'baseline' describes data measured prior to randomization. In this paper, we present a semiparametric locally efficient estimator to improve efficiency of inferences in randomized trials with correlated outcomes when baseline covariates are available. We begin with a review of current estimators for multivariate outcomes and then introduce our locally efficient estimator.

Cluster randomized and longitudinal trials, which are widespread in medical research, are two examples of randomized studies that have correlated outcomes. The outcome for the i_{th} independent randomized unit, $i = 1, ..., m,$ in such studies is denoted by the n_i-dimensional response vector $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^{\mathrm{T}}$, which may represent longitudinal measurements taken on the same individual or a set of responses from individuals within a common cluster such as a family, hospital, or class. Considering the substantial costs incurred by randomizing groups or following subjects over time, it is of interest to determine how to most efficiently estimate treatment effects using all available data.

Generally, i.i.d. data $O_i = (\mathbf{Y}_i, A_i, \mathbf{X}_i)$ are observed, where A_i denotes a scalar treatment assignment for 1 of K possible treatments, and $\mathbf{X}_i$ is a matrix of baseline covariates. Throughout we allow n_i to be random but assume it is ignorable. Longitudinal data also consist of a time variable $t_i = (t_{i1}, t_{i2}, ..., t_{in_i})^{\mathrm{T}}$ denoting time points at which outcomes are measured. As in the case of unit size n_i, we allow t_i to be random but assume it is ignorable. When repeated measures are taken on the same subject, baseline covariates are measured at $t_{ij} = 0$; thus $X_{ij} = X_i$ for all $j = 1, 2, ..., n_i$, resulting in a single level of baseline covariate information. Clustered data, however, may include pre-treatment covariates at the level of the group or the individual, creating two layers of auxiliary data. In the longitudinal context, we refer to the vector $\mathbf{Y}_i$ as the subject, or independent unit

and Y_{ij} as observation- or measurement-level data. For clustered data, we refer to $\mathbf{Y}_i$ as cluster-level and Y_{ij} as individual-level.

Classical approaches to estimation of marginal effects often involve specifying a restricted mean model for expected outcomes given treatment assignment, and therefore only use data on treatment and outcome in estimation. Specifically, in longitudinal studies, the marginal effect of treatment over time may be measured by assuming the restricted mean model

$$E(Y_{ij}|A_i, t_{ij}) = g\{\beta_0 + \beta_A A_i + \beta_t^{\mathrm{T}} f_1(t_{ij}) + \beta_{A,t}^{\mathrm{T}} A_i f_1(t_{ij})\}, \tag{2.1}$$

where $f_1(t_{ij})$ is a choice of function of t_i. The main effect β_A, which measures imbalance in $E(Y_{ij}|A_i, t_{ij})$ at baseline, is expected to be zero when randomization successfully balances covariate profiles across treatment arms. The post-baseline effect of treatment is measured by $\beta_{A,t}$. Parameters β_t and $\beta_{A,t}$ may be multivariate, as the effect of time on expected outcomes may be of some polynomial form. Similarly, for clustered data, the semiparametric model

$$E(Y_{ij}|A_i) = g(\beta_0 + \beta_1 A_i) \tag{2.2}$$

may be assumed, with treatment effects determined by inference on β_1.

Semiparametric estimates of treatment effects may be obtained by solving generalized estimating equations (GEE), as introduced by Liang and Zeger (1986). Given model (2.1) and data $W_i = (\mathbf{Y}_i, A_i)$, the p-dimensional coefficient vector $\beta = (\beta_0, \beta_1, ..., \beta_p)$ may be estimated by solving the GEE

$$\sum_{i=1}^{m} \psi(W_i; h, \beta) = \sum_{i=1}^{m} h(A_i, t_i)\{\mathbf{Y}_i - \mathbf{g}(A_i, t_i; \beta)\} = \mathbf{0}, \tag{2.3}$$

where the index or weight $h(A_i, t_i)$ is a $p \times n_i$ function of a random treatment variable A_i and time t_i, and $\mathbf{g}(A_i, t_i; \beta) = \{g(A_i, t_0; \beta), g(A_i, t_1; \beta), ..., g(A_i, t_{n_i}; \beta)\}^{\mathrm{T}}$. We use bold $\mathbf{g}(A_i, t_i; \beta)$ to denote the vector-valued mean function and $g(A_i, t_{ij}; \beta)$ to represent its scalar components. In semiparametric theory, the set $\{\psi(h; W_i) : h\}$, indexed by h, is derived as the orthogonal complement of the nuisance tangent space of model (2.1), $\Lambda_{nuis}^{\perp}$,

where Λ_{nuis} is defined as the closure of the linear span of all nuisance scores corresponding to smooth parametric submodels {Bickel et al. (1993)}. Here a nuisance parameter t under a smooth parametric submodel $F_t(O)$ satisfies $\frac{\partial \beta(F_t)}{\partial t} = 0$, where $\beta(F_t)$ is the parameter of interest under F_t, and $F_0 = F$, the data generating mechanism. The orthogonal complement, $\Lambda_{nuis}^{\perp}$, contains the set of all estimating functions of β {Bickel et al. (1993); van der Vaart (1998)}. When no baseline covariates are observed, Chamberlain (1986) shows that the efficient score of β, is obtained by setting $h(A_i, t_i) = \mathbf{D}_i^{\mathrm{T}} \mathbf{V}_i^{-1}$, where $\mathbf{V}_i$ is the $n_i \times n_i$ variance-covariance matrix of $\mathbf{Y}_i$, and $\mathbf{D}_i = \frac{\partial \mathbf{g}(A_i, t_i; \beta)}{\partial \beta}$.

Considering data O_i, which contain covariates $\mathbf{X}_i$, recent developments have resulted in a class of estimators that improve efficiency by augmenting standard estimating equations. When baseline covariates are predictive of the outcome these estimators reduce variability in estimated treatment effects, irrespective of the outcome distribution. Augmented estimators are constructed by starting with a standard estimating function and subtracting the orthogonal projection of the standard estimating function onto the span of the scores of the treatment mechanism {Robins et al. (1994), Robins (1999)}. For correlated outcomes, $\Lambda_{nuis}^{\perp} = \{\psi(O_i, h, \gamma, \beta) : h, \gamma\}$, and augmented GEE are

$$\sum_{i=1}^{m} \psi_i(O_i; \beta, h, \gamma) = \sum_{i=1}^{m} \left[\overbrace{h(A_i, t)\{\mathbf{Y}_i - \mathbf{g}(A_i, t_i; \beta)\}}^{Standard\ GEE} - \overbrace{\sum_{k=0}^{K-1} \{I(A_i = k) - \pi_k\} \gamma_k(\mathbf{X}_i)}^{arbitrary\ score\ of\ [A|X]} \right] = \mathbf{0}, \tag{2.4}$$

where for K-level treatment A_i, $P(A_i = k) = \pi_k$. Fixing $h(A_i, t_i)$, the most efficient estimating function sets $\gamma_k(\mathbf{X}_i) = \gamma_{k_{opt}}(\mathbf{X}_i) = h(k, t)\{E(Y|A_i = k, \mathbf{X}_i, t) - \mathbf{g}(k, t; \beta)\}$ {Robins et al. (1994), Robins (2000); van der Laan and Robins (2003); Zhang et al. (2008)}. The augmentation therefore involves estimation of the conditional mean outcome regression model $E(\mathbf{Y}_i|\mathbf{X}_i, A_i)$. Recalling longitudinal marginal model (2.1), if outcomes Y_{ij} are restricted to post-baseline measurements, the baseline measurement Y_{i0} may be utilized as a baseline covariate and included in $\mathbf{X}_i$. The interpretation of model parameters then changes, with the effect of treatment over time evaluated through β_A and $\beta_{A,t}$. In contrast to the previous interpretation, β_A now measures a constant shift in $g^{-1}\{E(Y_{ij}|A_{ij}, t_{ij})\}$

due to treatment, while nonzero $\beta_{A,t}$ indicates a change in the impact of treatment on $g^{-1}\{E(Y_{ij}|A_{ij}, t_{ij})\}$ over time.

Given a semiparametric model, a locally efficient estimator is defined as an estimator that achieves the semiparametric efficiency bound at a given submodel for the data-generating law, but remains consistent outside of the data-generating submodel {Bickel et al. (1993)}. Locally efficient estimators of parameters in restricted mean models of marginal treatment effects have been implemented for univariate data in the presence of baseline covariates by Robins (2000); Bang and Robins (2005); van der Laan and Rubin (2006); Tsiatis et al. (2008); Zhang et al. (2008); Moore and van der Laan (2009b), and Moore and van der Laan (2009a). For a univariate outcome, the model $g_s(A_i; \beta)$ containing a unique parameter for each treatment level is saturated. Under a saturated model, the choice of the index function $h(\cdot)$ has no impact on the resulting asymptotic variance and is therefore not considered for deriving efficient estimators. When $\mathbf{Y}_i$ is multivariate, $g_s(A_i; \beta)$ is not saturated, as the saturated model would allow different mean models for each element of the vector. As a result, $\Lambda_{nuis}^{\perp}$ provides a larger set of estimating functions indexed by $h(\cdot)$, where the choice of $h(\cdot)$ impacts efficiency. One particular index function, referred to as the optimal index, defines the efficient score and corresponding locally efficient estimator.

Robins (1999) established general theory for deriving locally efficient estimators of treatment effects in marginal structural models (MSMs) of time-dependent exposures, including the case of multivariate outcomes. These estimators, however, were not implemented nor evaluated in practice. Models (2.1) and (2.2) may be viewed as examples of MSMs for a point exposure; the Robins (1999) theory therefore equally applies. The locally efficient augmented estimator does not generally have the same optimal index $h(A_i, t_i)$ as the standard, unaugmented estimator. When incorporating auxiliary covariates in the estimation of marginal treatment effects via augmented GEE, the choice $h(A_i, t_i) = \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}$, while resulting in a consistent estimator, is therefore no longer optimal. The semiparametric efficient estimator is determined by optimizing over all $p \times n_i$ index functions

$h(A_i, t_i)$ {Robins et al. (1994); Robins (1999); van der Laan and Robins (2003)}. Although semiparametric efficient estimators may be obtained theoretically, they are often computationally intensive to calculate. Consequently, inefficient estimators are typically used. The suboptimal estimator based on augmenting GEE with the standard index function was shown to improve efficiency by Stephens et al. (2012a). In subsequent text, we refer to unaugmented GEE (2.3) with the index function $h(A_i, t_i) = \mathbf{D}_i^{\mathrm{T}} \mathbf{V}_i^{-1}$ as Standard GEE, and the suboptimal estimator obtained by augmenting Standard GEE is referred to as Simple Augmented GEE. Here we further improve on Simple Augmented GEE by deriving the corresponding semiparametric locally efficient estimator.

The following section presents the locally efficient estimator of marginal treatment effects in randomized trials with correlated outcomes when auxiliary data are available. Our estimator builds upon Standard GEE using principles from semiparametric theory. We construct these estimators by deriving a closed form expression of the efficient score for a variety of restricted mean models for the marginal treatment effect. We also discuss an implementation procedure detailing how to appropriately estimate each component of the efficient score. In Sections 2.3 and 2.4 we compare the derived semiparametric locally efficient estimator to standard and Simple Augmented GEE through simulations and application to the AIDS Clinical Trial Group study 398, a randomized longitudinal HIV intervention trial.

2.2 Methods

2.2.1 The Efficient Score

We consider the setting of longitudinal data and note that results follow analogously for clustered data by omitting t_i. Before presenting the main result, some additional notation is required. Conditioning on t_i, the matrix $h(A_i, t_i)$ takes K possible values, which may be denoted by K $p \times n_i$ constant matrices $h_0(t_i),\ h_1(t_i), ...,\ h_{K-1}(t_i)$. For binary treatment, we have $\mathbf{h}_1 = h_1(t_i)$ and $\mathbf{h}_0 = h_0(t_i)$, which denote the index func-

tions under treatment $(A = 1)$ and control $(A = 0)$, respectively. Let $\Delta_{k_i}(X) = E(\mathbf{Y}_i|A_i = k, \mathbf{X}_i, t_i) - \mathbf{g}(k, t_i; \beta)$, the n_i-dimensional vector of the difference in the conditional and marginal mean outcomes given time. Using this construction, let $\mathbf{h} = [\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{K-1}]$, the complete index matrix of dimension $p \times Kn_i$. Using a result from Newey and McFadden (1994), we show in Appendix D that the optimal index $h_{opt}(A, t)$ and resulting efficient score may be determined by solving a generalized information equality. Here we present our main result:

Proposition 1. *The efficient score for model* (2.1) *given data* $O = (\mathbf{Y}, A, \mathbf{X})$ *is*

$$\mathbf{h_{opt}} = \left[\pi_0 \frac{\partial \mathbf{g}(0, t; \beta)}{\partial \beta^{\mathrm{T}}}, \pi_1 \frac{\partial \mathbf{g}(1, t; \beta)}{\partial \beta^{\mathrm{T}}}, ..., \pi_{K-1} \frac{\partial \mathbf{g}(K-1, t; \beta)}{\partial \beta^{\mathrm{T}}}\right]^{\mathrm{T}} \mathbf{C}^{-}, \tag{2.5}$$

$\mathbf{C} = \mathbf{C}_1 - \mathbf{C}_2$, where

$$\mathbf{C}_1 = \begin{bmatrix} \pi_0 V(\mathbf{Y}|A=0) & 0 & \cdots & 0 \\ 0 & \pi_1 V(\mathbf{Y}|A=1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_{K-1} V(\mathbf{Y}|A=K-1) \end{bmatrix},$$

and

$$\mathbf{C_2} = \begin{bmatrix} \pi_0(1-\pi_0) E_X\left[\Delta_0(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X})\right] & \cdots & -\pi_0\pi_{K-1} E_X\left[\Delta_0(\mathbf{X})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X})\right] \\ -\pi_1\pi_0 E_X\left[\Delta_1(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X})\right] & \ddots & -\pi_1\pi_{K-1} E_X\left[\Delta_1(\mathbf{X})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X})\right] \\ \vdots & \ddots & \vdots \\ -\pi_{K-1}\pi_0 E_X\left[\Delta_{K-1}(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X})\right] & \cdots & \pi_{K-1}(1-\pi_{K-1}) E_X\left[\Delta_{K-1}(\mathbf{X})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X})\right] \end{bmatrix}.$$

As shown above, $\mathbf{C}$ is of dimension $Kn_i \times Kn_i$ and may be decomposed into the difference $\mathbf{C} = \mathbf{C}_1 - \mathbf{C}_2$, where $\mathbf{C}_1$ is a block diagonal matrix with diagonal components $\pi_k V(Y|A = k, t)$. The block diagonal of $\mathbf{C}_2$ contains the matrices $\pi_k(1 - \pi_k) E_X\left[\Delta_k(X)\Delta_k^{\mathrm{T}}(X)\right]$, and off-diagonal block components are determined by $-\pi_k\pi_{k'} E_X\left[\Delta_k(X)\Delta_{k'}^{\mathrm{T}}(X)\right]$.

When treatment is binary, $\mathbf{C}$ simplifies to

$$\mathbf{C} = \begin{bmatrix} \pi_1 V(Y|A=1,t) - \pi_1\pi_0\zeta^{1,1} & \pi_1\pi_0\zeta^{1,0} \\ \pi_1\pi_0\zeta^{0,1} & \pi_0 V(Y|A=0,t) - \pi_1\pi_0\zeta^{0,0} \end{bmatrix}, \tag{2.6}$$

where $\pi_0 = 1 - \pi_1$, and $\zeta^{a,a'} = E_X\left[\Delta_a(X)\Delta_{a'}^{\mathrm{T}}(X)\right]$. Inverting $\mathbf{C}$ analytically,

$$
\begin{aligned}
h_{opt}(A) = & \Bigg\{ D^{\mathrm{T}}(A) - \pi_1^{1-A}(1-\pi_1)^A D^{\mathrm{T}}(1-A)\times \\
& \left[V(1-A) - \pi_1^A(1-\pi_1)^{1-A}\zeta^{1-A,1-A}\right]^{\mathrm{T}^{-1}} \zeta^{1-A,A} \Bigg\} \times \\
& \Bigg\{ V(A) - \pi_1^{1-A}(1-\pi_1)^A \times \\
& \left(\zeta^{A,A} + \zeta^{A,1-A}\left[\frac{V(1-A)}{\pi_1^A(1-\pi_1)^{1-A}} - \zeta^{1-A,1-A}\right]^{-1} \zeta^{A,1-A^{\mathrm{T}}} \right) \Bigg\}^{-}. \qquad (2.7)
\end{aligned}
$$

Expressing the optimal index as in (2.7), it is clear that $\mathbf{h_{opt}}$ incorporates information on the treatment assignment and auxiliary covariates $\mathbf{X}$ through $\zeta^{a,a'}$, while the standard index $h_{std} = D^{\mathrm{T}}(A)V(A)^{-1}$, does not. The matrix $\zeta^{a,a'}$ is by definition the covariance of $E(\mathbf{Y}_i|\mathbf{X}_i, a, t_i)$ and $E(\mathbf{Y}_i|\mathbf{X}_i, a', t_i)$, the expected outcomes given baseline covariates and treatment assignment to a and a', respectively. The optimal index $\mathbf{h_{opt}}$ therefore boosts efficiency by incorporating information on the covariance in expected outcomes when weighting the residuals $\mathbf{Y}_i - \mathbf{g}(A_i;\beta)$ in the marginal model estimating equations. To implement Locally Efficient GEE, estimates of $V(\mathbf{Y}_i|A_i, t_i)$, $E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]$, and $\zeta^{a,a'}$ for all unique pairs of treatment levels $\{a, a'\}$, including $a = a'$, are needed. The next section details an estimation procedure for each component of $\mathbf{h_{opt}}$ when $\mathbf{Y}_i$ is continuous and $g(\cdot)$ is the identity link, or $\mathbf{Y}_i$ is binary and $g(\cdot)$ is the inverse logit link.

2.2.2 Estimation of $\mathbf{h_{opt}}$

The semiparametric locally efficient estimator requires estimates of $E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]$, $\zeta^{a,a'} = Cov\{E(\mathbf{Y}_i|\mathbf{X}_i, A_i = a, t_i), E(\mathbf{Y}_i|\mathbf{X}_i, A_i = a', t_i)|A_i, t_i\}$, and $V(\mathbf{Y}_i|A_i, t_i)$. These quantities may be linked by the law of total variance, $V(\mathbf{Y}_i|A_i, t_i) = E[V(\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i)|A_i, t_i] + V(E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]|A_i, t_i)$. For the i_{th} independent unit, the n_i-dimensional vector $E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]$ determines the $n_i \times n_i$ matrix $V(E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]|A_i, t_i)$ and ultimately impacts the form of the marginal variance matrix $V(\mathbf{Y}_i|A_i, t_i)$. Observing the relationship among each of these parameters provides guidance for model selection. For ex-

ample, the working marginal covariance selected must be compatible with the working model chosen for $E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]$. More generally, the models for each component of $\mathbf{h_{opt}}$ must be specified so that the model selected for one component does not preclude the model chosen for another. One approach that ensures compatibility is to first estimate $E(Y_{ij}|X_{ij}, A_i, t_{ij})$ through an appropriate regression technique to provide an estimate $\hat{E}(\mathbf{Y}_i|\mathbf{X}_i, A_i = a, t_i)$. The conditional mean outcome may be modeled by

$$E[Y_{ij}|X_{ij}, A_i, t_{ij}] = g\{\eta_0 + \eta_A A_i + \eta_t^{\mathrm{T}} f(t_{ij}) + \eta_{A,t}^{\mathrm{T}} A_i f(t_{ij}) + \eta_X^{\mathrm{T}} X_{ij} + \eta_{X,t}^{\mathrm{T}} X_{ij} f(t_{ij}) + \eta_{A,X}^{\mathrm{T}} A_i X_{ij}\}, \tag{2.8}$$

where X_{ij} respresents the collection of covariates for the j_{th} measurement in the i_{th} unit. Second, the covariance of the conditional expectation is estimated by noting how the model of $E(Y_{ij}|X_{ij}, A_i = a, t_i)$ impacts the form of the matrix $\zeta^{a,a'}$. Finally, $V(\mathbf{Y}_i|A_i, t_i)$ is estimated incorporating the estimates of $E[V(\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i)|A_i, t_i]$ and $V(E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]|A_i, t_i)$.

General estimation of $\zeta^{a,a'}$ and $V(\mathbf{Y}|A)$

Generally, $\zeta^{a,a'}$ may be estimated in a similar fashion to estimates of the correlation parameters in Standard GEE. Let $\zeta^{a,a'} = \mathbf{R}^{1/2}\mathbf{S}\mathbf{R}^{1/2}$, where $\mathbf{R}$ is a $n_i \times n_i$ diagonal matrix with the j^{th} diagonal component $R_{j,j} = V(E[Y_{ij}|X_{ij}, A_i, t_{ij}]|A_i, t_{ij}) = \nu_j^{a,a'}$, and $\mathbf{S}$ is a $n_i \times n_i$ correlation matrix with $S_{j,j} = 1$ and $S_{j,j'} = f(\tau^{a,a'})$. Parameter $\tau^{a,a'}$, which may be multivariate, and $\nu^{a,a'} = (\nu_1^{a,a'}, \nu_2^{a,a'}, ..., \nu_{n_i}^{a,a'})$ characterize the covariance in conditional mean outcomes under treatments a and a'. Letting $\hat{\Delta}_{a_{ij}} = \hat{E}(Y_{ij}|X_{ij}, A_i = a, t_{ij}) - g(a, t_{ij}; \hat{\beta}_{init})$, where $\hat{\beta}_{init}$ is an initial estimate of β, $\nu_j^{a,a'}$ may be estimated by

$$\hat{\nu}_j^{a,a'} = \frac{1}{m - p_\eta} \sum_{i=1}^{m} \hat{\Delta}_{a_{ij}} \hat{\Delta}_{a'_{ij}}, \tag{2.9}$$

where p_η is the dimension of the outcome regression parameter η. The correlation parameter $\tau^{a,a'}$ is then estimated by the moment equations

$$\sum_{i=1}^{m} \sum_{j<j'} \left\{ \frac{\hat{\Delta}_{a_{ij}}}{\sqrt{\hat{\nu}_j^{a,a'}}} \frac{\hat{\Delta}_{a'_{ij'}}}{\sqrt{\hat{\nu}_{j'}^{a,a'}}} - f(\tau^{a,a'}) \right\} = 0. \tag{2.10}$$

For $a = a'$, we obtain an estimate of $\zeta^{a,a} = V(E[\mathbf{Y}_i|\mathbf{X}_i, A_i = a, t_i])$.

As an alternative approach, one may also derive a complex expression of $\zeta^{a,a'}$ that depends on $\eta = (\eta_0, \eta_A, \eta_t^{\mathrm{T}}, \eta_{A,t}^{\mathrm{T}}, \eta_X^{\mathrm{T}}, \eta_{X,t}^{\mathrm{T}}, \eta_{A,X}^{\mathrm{T}})^{\mathrm{T}}$ and the covariance in baseline covariates. An empirical estimate of $Cov(\mathbf{X}_i)$ may then be substituted into this expression.

After estimating $\zeta^{a,a}$, the conditional variance of $\mathbf{Y}_i$, $V(\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i)$, may be estimated using the correlation parameters from GEE based on the conditional mean model (2.8). Under homoscedasticity $V(\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i) = \lambda$ for all i. The marginal variance $V(\mathbf{Y}_i|A_i, t_i)$ is then estimated by $\hat{V}(\mathbf{Y}_i|A_i, t_i) = \hat{\zeta}^{a,a} + \hat{\lambda}$, where $\hat{\zeta}^{a,a}$ and $\hat{\lambda}$ are estimates of $\zeta^{a,a}$ and λ, respectively.

Practical estimation of $\zeta^{a,a'}$

For clustered data and longitudinal data with $\eta_{X,t} = \mathbf{0}$ in (2.8), calculating $\zeta^{a,a'}$ is straightforward. When data are clustered, $\eta_t = \eta_{A,t} = \eta_{X,t} = \mathbf{0}$, leaving $E[Y_{ij}|X_{ij}, A_i] = g(\eta_0 + \eta_A A_i + \eta_X^{\mathrm{T}} X_{ij} + \eta_{A,X}^{\mathrm{T}} A_i X_{ij})$. In this setting, $\zeta_{j,j'}^{a,a'}$, the j, j' element of $\zeta^{a,a'}$, is calculated as $\zeta_{j,j'}^{a,a'} = Cov_X\{g(\eta_0 + \eta_A a + \eta_X^{\mathrm{T}} X_{ij} + \eta_{A,X}^{\mathrm{T}} a X_{ij}),\ g(\eta_0 + \eta_A a' + \eta_X^{\mathrm{T}} X_{ij'} + \eta_{A,X}^{\mathrm{T}} a' X_{ij'})\}$. If auxiliary covariates $X_{ij}, X_{ij'}$ are equally correlated among subjects within a cluster $\zeta_{j,j'}^{a,a'} = \rho_{a,a'}$ for all j, j'. This holds for all link functions $g(\cdot)$. For longitudinal data when $\eta_{X,t} = \mathbf{0}$ (i.e. the effects of baseline covariates on the conditional mean outcome do not vary over time) $\zeta_{j,j'}^{a,a'} = Cov_X\{g(\eta_0 + \eta_A a + \eta_t^{\mathrm{T}} f(t_{ij}) + \eta_{A,t}^{\mathrm{T}} a f(t_{ij}) + \eta_X^{\mathrm{T}} X_i + \eta_{A,X}^{\mathrm{T}} a X_i),\ g(\eta_0 + \eta_A a' + \eta_t^{\mathrm{T}} t_{ij'} + \eta_{A,t}^{\mathrm{T}} a' f(t_{ij'}) + \eta_X^{\mathrm{T}} X_i + \eta_{A,X}^{\mathrm{T}} a' X_i\}$. If $g(\cdot)$ is the identity link, this reduces to $\zeta_{j,j'}^{a,a'} = Cov(\eta_X^{\mathrm{T}} X_i + \eta_{A,X}^{\mathrm{T}} a X_i,\ \eta_X^{\mathrm{T}} X_i + \eta_{A,X}^{\mathrm{T}} a' X_i) = \rho_{a,a'}$ for all j, j', since $X_{ij} = X_i$ for all j.

Practical estimation of $V(Y_i|A_i = a)$

In some special cases where summing $E[V(\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i)|A_i, t_i]$ and $V(E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]|A_i, t_i)$ results in a marginal covariance matrix $V(\mathbf{Y}_i|A_i, t_i)$ with a standard form, e.g., exchangeable, $V(\mathbf{Y}_i|A_i, t_i)$ may be estimated directly while maintaining compatibility with $E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]$. As stated above, if individual-level covariates X_{ij}

are equally correlated among subjects within the i_{th} cluster, the model $E[Y_{ij}|X_{ij}, A_i = a]$ imposes compound symmetry on $\zeta^{a,a'}$, where diagonal components depend on $Var(X_{ij})$ and off-diagonal components are determined by $Cov(X_{ij}, X_{ij'})$. If the conditional variance $V(\mathbf{Y}_i|\mathbf{X}_i, A_i)$ is also exchangeable, $V(\mathbf{Y}_i|A, t_i)$ will have an exchangeable structure. The optimal index $\mathbf{h_{opt}}$ may then be calculated by estimating $V(\mathbf{Y}_i|A_i, t_i)$ directly as in Standard or Simple Augmented GEE and using the above procedure to estimate $\zeta^{a,a'}$.

A consistent estimator of the asymptotic variance of $\hat{\beta}_{opt}$, the solution to the augmented estimating equations (2.4) evaluated under (2.5), may be calculated using the sandwich variance formula of Huber (1964).

2.3 Simulation Study

Semiparametric Locally Efficient GEE were compared to Standard and Simple Augmented GEE through a simulation study. Simulations were completed for clustered data with continuous and binary outcomes and longitudinal data with continuous outcomes. Results are based on 1,000 Monte Carlo datasets.

2.3.1 Continuous Outcomes

Clustered Data

Data for $m = 500$ clusters were generated, with n_i=2,4,6,8,10,12 with equal probability for the first set of simulations and n_i=10,20,30,40,50 in the second set. Auxiliary covariates X_{ij1}, X_{ij2}, and X_{ij3} were each generated from a multivariate normal distribution with $Var(X_{ij1})$=2, $Var(X_{ij2})$=6, and $Var(X_{ij3})$=5. Correlation was induced among individual-level covariates within the same cluster by setting $Cov(X_{ij1}, X_{ij'1}) = \varsigma_{X_1}$, $Cov(X_{ij2}, X_{ij'2}) = \varsigma_{X_2}$, and $Cov(X_{ij3}, X_{ij'3})$=1. Covariance terms ς_{X_1} and ς_{X_2} were varied from 0.5 to 2 and 1.5 to 6, respectively, to evaluate the effect of auxiliary covariate correlation on the performance of Locally Efficient Augmented GEE. At ς_{X_1}=0.5 and ς_{X_2}=1.5

covariates were weakly correlated among individuals in the same cluster, while at ς_{X_1}=5 and $\varsigma_{X_{2,4}}$=6, covariates were perfectly correlated, thereby becoming cluster-level. The exact values considered for ς_{X_1} and ς_{X_2} were (0.5, 1, 1.5, 2) and (1.5, 3, 4.5, 6), for simulation sets 1-4 at each set of cluster sizes. Within the j_{th} individual in the i_{th} cluster, auxiliary covariates were independent. The treatment variable A_i was drawn from the Bernoulli distribution with p=1/2. Clustered responses were generated from the following model, with individual-level error terms $\varepsilon_{ij} \sim N(0, 40)$ and cluster-level effects $b_i \sim N(0, \sigma_b^2)$: $Y_{ij}|A_i, X_{ij}, b_i = 1.0 + 1.1X_{ij1}^2 + 0.9X_{ij2} + 0.5A_i + b_i + \varepsilon_{ij}$. The proportion of variability in Y_{ij} explained by auxiliary covariates X_{ij} was held fixed at roughly 25%. Simulations were completed with $\sigma_b^2 = 0$ and $\sigma_b^2 = 6$, representing the case in which covariates account for all between-cluster heterogeneity and the alternative of some intracluster correlation caused by an unmeasured variable, respectively.

For each dataset, the marginal effect of treatment was estimated by fitting model (2.2) through Standard, Simple Augmented, and Locally Efficient Augmented GEE. The impact of misspecification on the locally efficient estimator and its efficiency relative to Simple Augmented and Standard GEE was evaluated by fitting various models to estimate $E(\mathbf{Y}|\mathbf{X}, A)$. The correct model for $E(\mathbf{Y}|\mathbf{X}, A)$, denoted by 'C' in tables and figures, was $E(Y_{ij}|X_{ij}, A_i) = \eta_0 + \eta_1 X_{ij1}^2 + \eta_2 X_{ij2} + \eta_3 A_i$, and two incorrect models were Wrong 1, 'W1'=$E(Y_{ij}|X_{ij}, A_i) = \eta_0 + \eta_1 X_{ij1} + \eta_2 X_{ij2} + \eta_3 A_i$ and Wrong 2, 'W2'=$E(Y_{ij}|X_{ij}, A_i) = \eta_0 + \eta_1 X_{ij1}^2 + \eta_2 X_{ij2} + \eta_3 X_{ij3} + \eta_4 A_i$. 'Wrong 1' evaluated the impact of misspecifying the functional form of X_{ij1}, while 'Wrong 2' examined the effect of adding noise to the outcome regression model. All working covariance matrices were fit under exchangeable structure.

Large cluster efficiency comparisons relative to Standard GEE are summarized in Figure 2.1, while the Monte Carlo Relative Efficiency (MCRE) of the locally efficient estimator to Simple Augmented GEE may be found in Tables 2.1(a)-2.1(b). Small cluster results are presented in Figure 2.2. Across all levels of correlation, augmented estimators resulted in increased efficiency compared to the unaugmented estimator (MCRE 1.25-

Table 2.1: **Monte Carlo Relative Efficiency of Locally Efficient Augmented GEE to Sub-optimal Augmented GEE:** Continuous clustered outcomes. Working Marginal Covariance (WMCov): Exchangeable (Exch). Outcome Regression (OR): Correct (C), Wrong 1(W1), Wrong 2 (W2). First entry $\sigma_b^2 = 0$, second entry $\sigma_b^2 = 6$. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X,A)\}$.

(a) Cluster Size = 2,4,6,8,10,12

	Correlation among X_{ij}			
WMCov/OR	0.25	0.50	0.75	1.00
Exch/C	1.0115	1.0450	1.0907	1.1464
	1.0036	0.9991	1.0010	1.0085
Exch/W1	1.0062	1.0089	1.0064	1.0038
	1.0006	1.0008	1.0018	1.0019
Exch/W2	1.0114	1.0448	1.0905	1.1462
	1.0036	0.9990	1.0009	1.0083

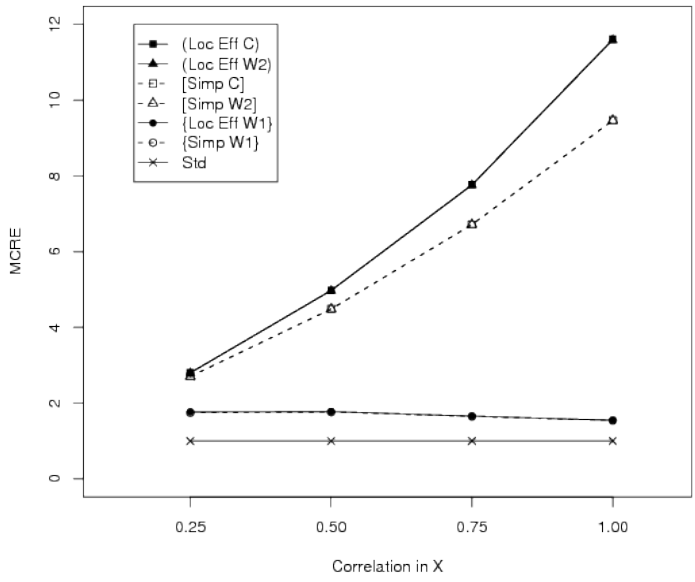
(b) Cluster Size = 10,20,30,40,50

	Correlation among X_{ij}			
Cov/OR	0.25	0.50	0.75	1.00
Exch C	1.0356	1.1096	1.1563	1.2259
	1.0005	0.9999	1.0002	1.0011
Exch W1	1.0126	1.0081	1.0050	1.0032
	1.0000	1.0000	1.0001	1.0003
Exch W2	1.0352	1.1090	1.1556	1.2247
	1.0006	0.9998	1.0001	1.0009

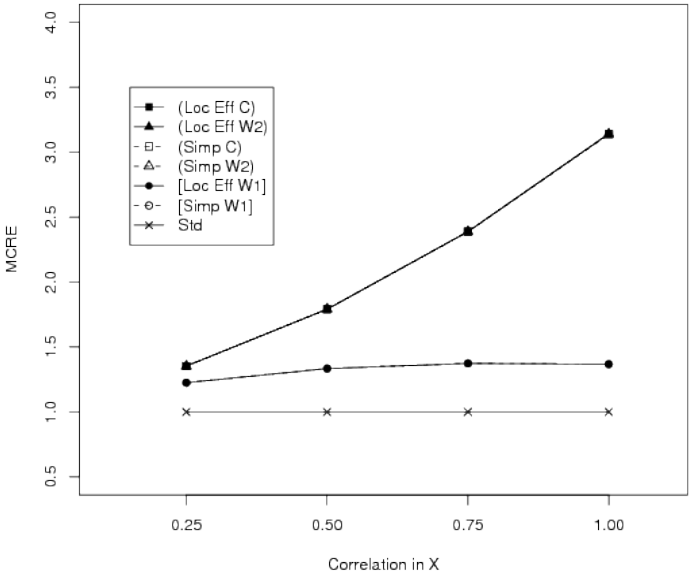
11.6). For low correlation among X_{ij} simple augmented and locally efficient estimators performed similarly. When correlation was increased among X_{ij} within a cluster, Locally Efficient GEE gained in efficiency over Simple Augmented GEE (MCRE Locally Efficient to Simple Augmented GEE 1.01-1.22). Increased covariance among auxiliary covariates also resulted in greater efficiency gains for any augmented GEE relative to the standard estimator. Trends were more pronounced for large average cluster size (average n_i=30 vs. average n_i=7).

Figure 2.1: **MCRE of Locally Efficient and Simple Augmented GEE Relative to Standard (Unaugmented) GEE:** Continuous clustered outcomes, large clusters. Estimators corresponding to each curve are denoted by 'Estimator-Outcome Regression' using the abbreviations: Loc Eff-Locally Efficient, Simp-Simple Augmented, Std-Standard; C-Correct, W1-Wrong 1, W2-Wrong 2. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X,A)\}$. The order of curves in the legend follows the order of curves on the figure, with sets of superimposed curves denoted by '()', '[]', or'{}'.

Figure 2.1(Continued)



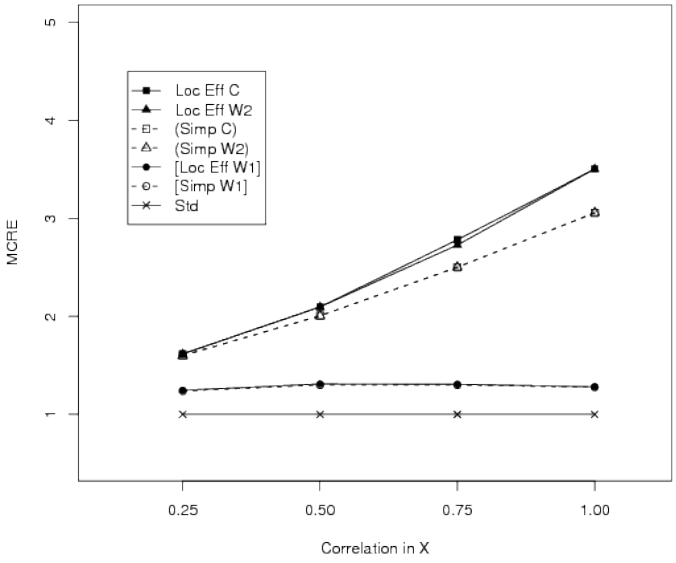
(a) n_i=(10,20,30,40,50), $\sigma_b^2 = 0$



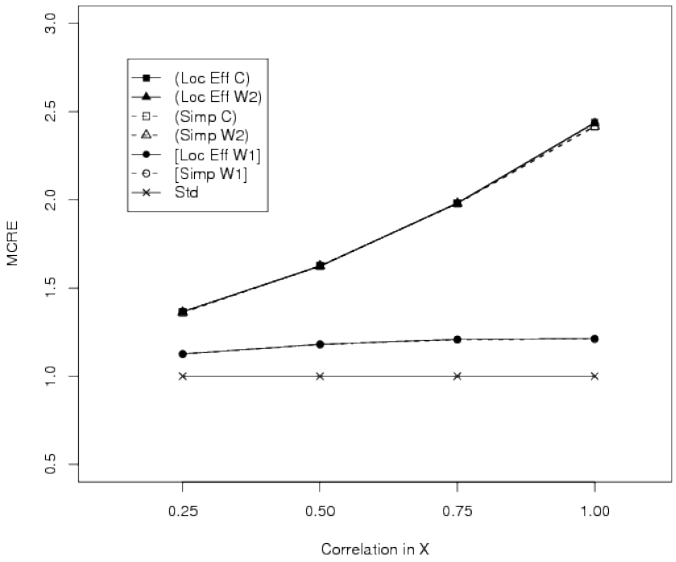
(b) n_i=(10,20,30,40,50), $\sigma_b^2 = 6$

Figure 2.2: **MCRE of Locally Efficient and Simple Augmented GEE Relative to Standard (Unaugmented) GEE:** Continuous clustered outcomes, small clusters. Estimators corresponding to each curve are denoted by 'Estimator-Outcome Regression' using the abbreviations: Loc Eff-Locally Efficient, Simp-Simple Augmented, Std-Standard; C-Correct, W1-Wrong 1, W2-Wrong 2. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X,A)\}$. The order of curves in the legend follows the order of curves on the figure, with sets of superimposed curves denoted by '()' and '[]'.

Figure 2.2(Continued)



(a) n_i=(2,4,6,8,10,12), $\sigma_b^2 = 0$



(b) n_i=(2,4,6,8,10,12), $\sigma_b^2 = 6$

Longitudinal Responses

For each Monte Carlo dataset, m=500 longitudinal response vectors $\mathbf{Y}_i$ were generated from the model $Y_{ij} = 1.5 + 1.1X_{i1}^2 + 0.9X_{i2} + 1.0t_{ij} + 1.0A_i + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, and $Cov(\varepsilon_{ij}, \varepsilon_{ij'})$ had an AR-1 structure with correlation parameter $\alpha = 0.1, 0.3,$ or 0.5 for different sets of simulations. Covariates X_{i1} and X_{i2} were normally distributed with mean 0 and variance σ_{X1}^2 and σ_{X2}^2, respectively. Variance parameters σ_ε^2, σ_{X1}^2, and σ_{X2}^2 were varied so that baseline covariates accounted for 10-60% of the variability in $\mathbf{Y}|A$ in increments of 10%. Subjects were randomly assigned to treatment (A_i=1) with probability 1/2. For each subject $t_i = (t_{i1} = 1, t_{i2} = 2, ..., t_{in_i} = n_i)$, where n_i varied from 1 to 8, as might be the case in a longitudinal study with staggered entry.

Standard GEE, Simple Augmented GEE, and Locally Efficient Augmented GEE were applied to each Monte Carlo dataset to estimate marginal treatment effects. All GEE were fit based on the marginal mean model $E(Y_{ij}|A_i) = \beta_0 + \beta_1 A_i + \beta_2 t_{ij}$ with inferences on the treatment effect completed through β_1. Standard and Simple Augmented GEE were applied to each Monte Carlo dataset with AR-1, exchangeable, and true working covariance structures, with the true structure under the marginal model being a summation of AR-1 and exchangeable matrices as described in section 2.2. Locally efficient augmented GEE were fit under the true covariance structure and a misspecified marginal AR-1 working covariance. Baseline covariates were incorporated fitting several outcome regression models. We use 'C' to denote the correct model $E(Y_{ij}|X_{ij}, A_i, t_{ij}) = \eta_0 + \eta_1 A_i + \eta_2 t_{ij} + \eta_3 X_{i1}^2 + \eta_4 X_{i2}$, which corresponds to the true data generating mechanism; 'W1' indicates the model $E(Y_{ij}|X_i, A_i, t_{ij}) = \eta_0 + \eta_1 A_i + \eta_2 t_{ij} + \eta_3 X_{i1} + \eta_4 X_{i2}$, omitting the exponent on X_{i1}; and 'W2' is the model that includes a noisy covariate X_{i3}, such that $E(Y_{ij}|X_i, A_i, t_{ij}) = \eta_0 + \eta_1 A_i + \eta_2 t_{ij} + \eta_3 X_{i1}^2 + \eta_4 X_{i2} + \eta_5 X_{i3}$.

Efficiency comparisons are summarized in Figure 2.3 and Table 2.2. The difference in the performance of locally efficient versus Simple Augmented GEE increased with the percent variability explained by $\mathbf{X}_i$ (MCRE of Locally Efficient to Simple Augmented GEE 1.0-1.15). Similarly, efficiency gains from augmenting increased with variability in

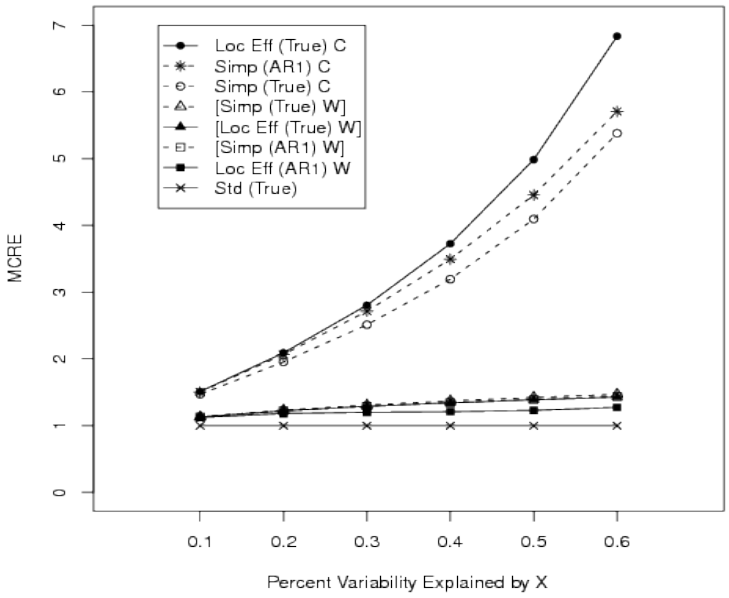
Table 2.2: **Monte Carlo Relative Efficiency of Locally Efficient Augmented GEE to Suboptimal Augmented GEE:** Continuous longitudinal outcomes. Working Marginal Covariance (WMCov): 1) True, exchangeable for $V(E(Y|X,A)|A)$ and AR1 for $V(Y|X,A)$ 2) AR1 for V(Y—A). Outcome Regression (OR): Correct (C), Wrong 1(W1), Wrong 2 (W2). First entry $\alpha = 0.1$, second entry $\alpha = 0.3$, third entry $\alpha = 0.5$.

	Correlation between Y and X					
WMCov/OR	10	20	30	40	50	60
True/C	1.0281	1.0700	1.1168	1.1662	1.2175	1.2702
	1.0166	1.0425	1.0728	1.1055	1.1398	1.1752
	1.0090	1.0234	1.0409	1.0603	1.0811	1.1028
True/W1	0.9995	0.9929	0.9851	0.9783	0.9735	0.9717
	1.0006	0.9974	0.9930	0.9887	0.9854	0.9837
	1.0009	0.9999	0.9982	0.9961	0.9943	0.9931
True/W2	1.0284	1.0703	1.1171	1.1664	1.2176	1.2701
	1.0168	1.0428	1.0731	1.1058	1.1401	1.1754
	1.0092	1.0237	1.0412	1.0606	1.0814	1.1031
AR1/W1	0.9916	0.9645	0.9300	0.8902	0.8832	0.8887
	0.9972	0.9858	0.9707	0.9567	0.9481	0.9481
	0.9996	0.9958	0.9903	0.9849	0.9811	0.9802

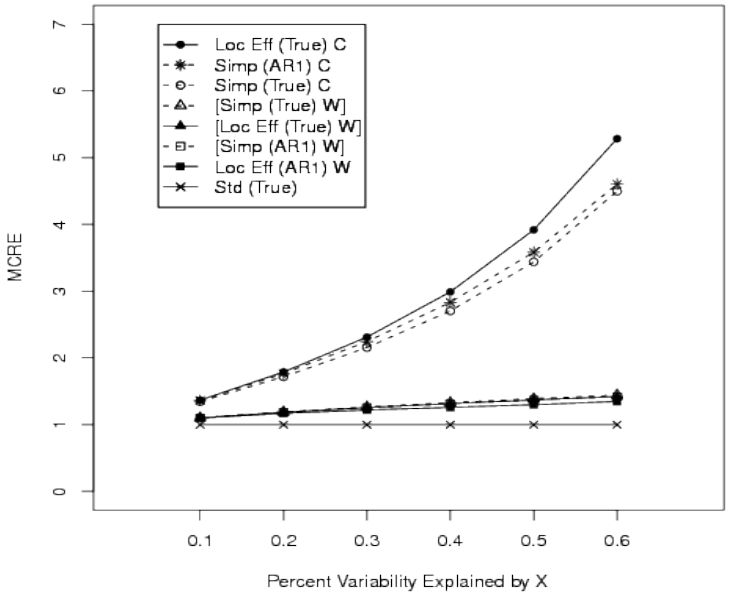
$\mathbf{X}_i$ (MCRE of Augmented GEE to Standard GEE 1.36-5.28). Among the simple augmented estimators, the estimator with the incorrect marginal AR-1 working covariance resulted in the β_1 estimate with the lowest variability. This illustrates an important distinction between locally efficient and suboptimal estimating functions. Among estimators using a suboptimal index, misspecified models for parameters in the index may result in more efficient inferences than correctly specified models. For the locally efficient estimator, asymptotic efficiency is achieved only in the absence of model misspecification for all parameters in the index function. It is also worthwhile to note that the simple augmented estimator with the marginal AR-1 covariance was slightly more efficient than the locally efficient GEE under the same misspecified marginal working covariance. This demonstrates that the locally efficent efficient GEE is a bit more sensitive to working marginal covariance misspecification than Simple Augmented GEE.

Figure 2.3: **MCRE of Locally Efficient and Simple Augmented GEE Relative to Standard (Unaugmented) GEE:** Continuous longitudinal outcomes. Estimators corresponding to each curve are denoted by 'Estimator (Marginal Working Covariance) Outcome Regression' using the abbreviations: Loc Eff-Locally Efficient, Simp-Simple Augmented, Std-Standard; AR1-Autoregressive(1) $V(Y|A)$, True-Exchangeable/AR1 for $V\{E(Y|X,A)\}$ and $V(Y|X,A)$, respectively; C-Correct, W1-Wrong 1, W2-Wrong 2;α=0.3. The order of curves in the legend follows the order of curves on the figure, with the set of superimposed curves denoted by '[]' and '{}'.

Figure 2.3(Continued)

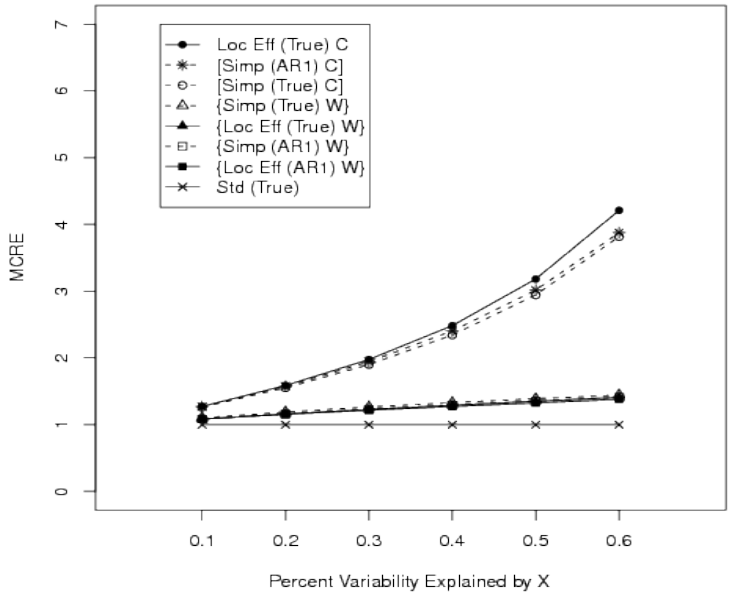


(a) α=0.1



(b) α=0.3

Figure 2.3(Continued)



(c) α=0.5

2.3.2 Clustered Binary Outcomes

As for continuous outcomes, data for m=500 clusters of variable size were generated with n_i=2,4,6,8,10,12 for small cluster settings and n_i=10,20,30,40,50 for the large cluster scenario. The binary treatment variable A_i was simulated from the Bernoulli(1/2) distribution. Individual-level covariates X_{ij1}, X_{ij2}, and X_{ij3} were each generated from a multivariate normal distribution with $\mu_{X_{ijk}} = 0$, $\sigma^2_{X_{ij1}} = \sigma^2_{X_{ij3}}$=2, $\sigma^2_{X_{ij2}} = 5$, and $Cov(X_{ijk}, X_{ij'k}) = \varsigma_{X_k}$, inducing marginal correlation among individuals within the same cluster. Covariance parameters ς_{X_k} were varied to evaluate the impact of covariance in auxiliary covariates on the performance of augmented estimators, with $\varsigma_{X_1} = \varsigma_{X_3} = 0.5, 1.0, 1.5, 2.0$ and $\varsigma_{X_3} = 1.25, 2.5, 3.75, 5.0$ for different sets of simulations. For low levels of ς_{X_k}, covariates were weakly correlated, while for $\varsigma_{X_k} = \sigma^2_{X_{ijk}}$, covariates were cluster-level. Binary outcomes were simulated from the model $logit[E(Y_{ij}|X_{ij}, A_i, b_i)] = 0.7{X_{ij1}}^2 + 0.4X_{ij2} - 0.5A_i + b_i$, where b_i was drawn from the bridge distribution for the logit link {Wang and Louis (2003)} with scale parameter θ. Simulations were completed

with two values of the bridge distribution scale parameter, $\theta = 1$ and $\theta = 0.8$, representing settings in which all sources of between-cluster heterogeneity are measured through auxiliary covariates, or when unmeasured sources of between-cluster heterogeneity are present. A total of 16 sets of simulations were done, varying cluster size, correlation in $\mathbf{X}$, and θ.

Standard, Simple Augmented, and Locally Efficient Augmented GEE were applied to each dataset and compared for efficiency. For each estimator, the model of interest was model (2.2) with $g(\cdot)$ the inverse logit link and β_1 measuring the marginal effect of treatment. Among augmented estimators, four outcome regression models were considered: 1) 'C'-Correct, $E(Y_{ij}|X_{ij}, A_i) = g(\eta_0 + \eta_1 X_{ij1}^2 + \eta_2 X_{ij2} + \eta_3 A_i)$; 2) 'W1'-Wrong 1, $E(Y_{ij}|X_{ij}, A_i) = g(\eta_0 + \eta_1 X_{ij1} + \eta_2 X_{ij2} + \eta_3 A_i)$; 3) 'W2'-Wrong 2, $E(Y_{ij}|X_{ij}, A_i) = g(\eta_0 + \eta_1 X_{ij1}^2 + \eta_2 X_{ij2} + \eta_3 X_{ij3} + \eta_4 A_i)$; and 4)'W1 OLS'-Wrong 1 OLS, $E(Y_{ij}|X_{ij}, A_i) = \eta_0 + \eta_1 X_{ij1} + \eta_2 X_{ij2} + \eta_3 X_{ij3} + \eta_4 A_i$. With the exception of model 4, which was fit through ordinary least squares (OLS), all outcome regression models were fit by logistic regression. All estimators were fit with exchangeable working covariances.

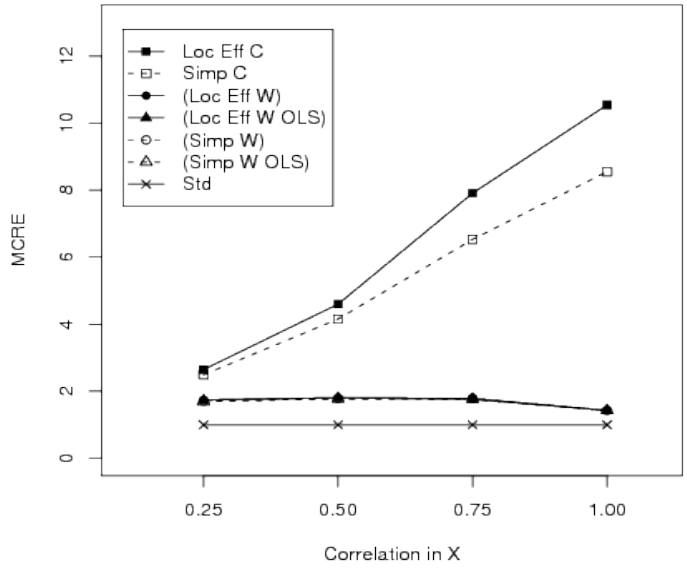
Large cluster results are shown in Figure 2.4 and Table 2.3. Figure 2.5 contains small cluster results. Conclusions are similar to those obtained for continuous outcomes. Efficiency improvement with augmented estimators relative to Standard GEE increased with correlation in auxiliary covariates (MCRE 1.10-10.54), as did the additional efficiency gains for the locally efficient GEE over Simple Augmented GEE (MCRE 1.0-1.23). Simple and locally efficient augmented estimators were equally efficient for $\theta = 0.8$, but differences in efficiency favoring the optimal estimator were observed for $\theta = 1$.

Table 2.3: **Monte Carlo Relative Efficiency of Locally Efficient Augmented GEE to Sub-optimal Augmented GEE:** Binary clustered outcomes. Working Marginal Covariance (WMCov): Exch-Exchangeable. Outcome Regression (OR): Correct (C), Wrong 1 (W1), Wrong 2 (W2), Wrong 1 Linear Model (W1-LM). First entry $\theta = 1$, second entry $\theta = 0.8$. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X, A)\}$.

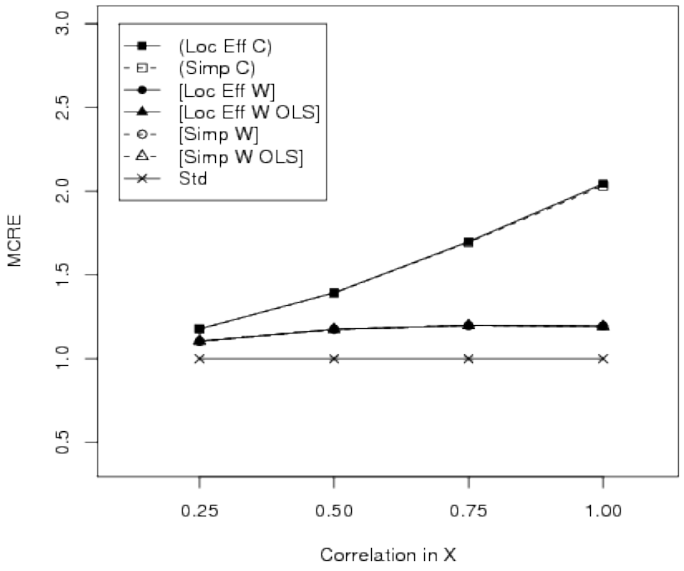
	Correlation between Y and X			
WMCov/OR	0.25	0.50	0.75	1.00
Exch/C	1.0624	1.1068	1.2113	1.2329
	0.9996	1.0009	1.0025	1.0057
Exch/W1	1.0247	1.0179	1.0025	1.0015
	1.0001	1.0003	1.0002	1.0001
Exch/W2	1.0630	1.1072	1.2080	1.2353
	0.9995	1.0009	1.0024	1.0056
Exch/W1-LM	1.0238	1.0171	1.0016	1.0008
	1.0001	1.0003	1.0001	1.0000

Figure 2.4: **MCRE of Locally Efficient and Simple Augmented GEE Relative to Standard (Unaugmented) GEE:** Binary clustered outcomes, large clusters. Estimators corresponding to each curve are denoted by 'Estimator-Outcome Regression' using the abbreviations: Loc Eff-Locally Efficient, Simp-Simple Augmented, Std-Standard; C-Correct, W1-Wrong 1, W2-Wrong 2. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X,A)\}$. The order of curves in the legend follows the order of curves on the figure, with sets of superimposed curves denoted by '()'and '[]'.

Figure 2.4(Continued)



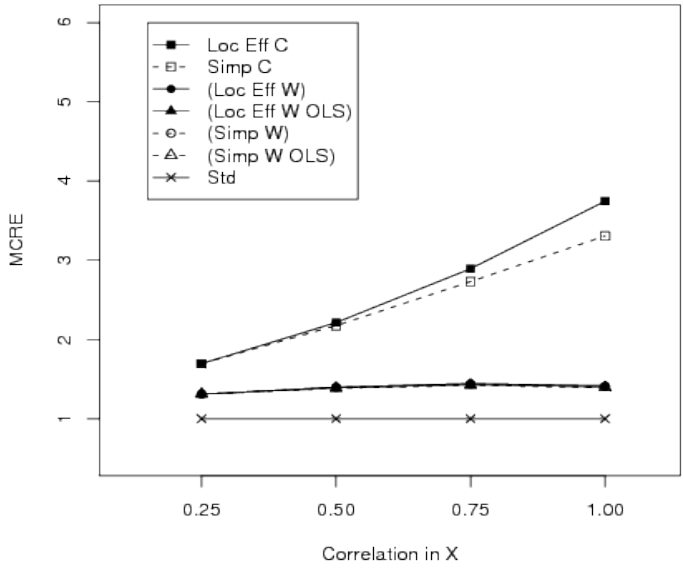
(a) n_i=(10,20,30,40,50), $\theta = 1$



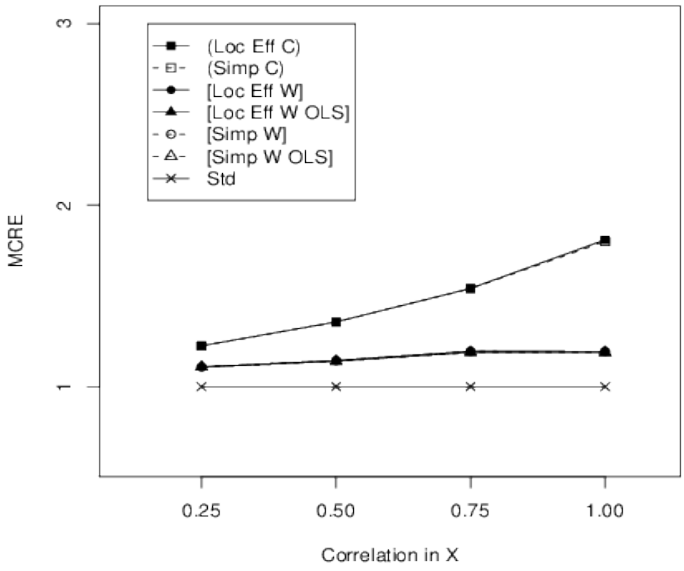
(b) n_i=(10,20,30,40,50), $\theta = 0.8$

Figure 2.5: **MCRE of Locally Efficient and Simple Augmented GEE Relative to Standard (Unaugmented) GEE:** Binary clustered outcomes, small clusters. Estimators corresponding to each curve are denoted by 'Estimator-Outcome Regression' using the abbreviations: Loc Eff-Locally Efficient, Simp-Simple Augmented, Std-Standard; C-Correct, W1-Wrong 1, W2-Wrong 2. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X,A)\}$. The order of curves in the legend follows the order of curves on the figure, with sets of superimposed curves denoted by '()'.

Figure 2.5(Continued)



(a) n_i=(2,4,6,8,10,12), $\theta = 1$



(b) n_i=(2,4,6,8,10,12), $\theta = 0.8$

2.4 Application: AIDS Clinical Trial Group Study 398

The semiparametric locally efficient estimator of marginal treatment effects for correlated outcomes was applied to data from AIDS Clinical Trial Group Study 398 (ACTG 398) {Hammer et al. (2002)}. ACTG 398 was a multicenter, double-blind trial, in which 481 HIV-infected patients were randomized to one of four arms, A) saquinavir, B) indinavir, C) nelfinavir, or D) placebo based on their past protease inhibitor (PI) treatment. Patients were only randomized to drugs to which they had no prior exposure. Randomized treatments were given to all participants in combination with antiretroviral therapy. Subjects' CD4 was measured at weeks 0 (baseline), 4, 8, and every 8 weeks thereafter until 48 weeks or dropout. GEE estimators were applied to compare the nelfinavir and placebo arms among patients who were eligible for both according to the stratified randomization scheme. Additional baseline covariates were age, sex, past PI use, past non-nucleoside reverse transcriptase inhibitor (NNRTI) exposure, weight, Karnofsky score, intravenous drug use, and race/ethnicity. Weeks 4-32 of followup were included for analysis, with CD4 measurements at week 4 and beyond included as outcomes and week 0 CD4 included as a baseline covariate. Data were approximately 90% complete through week 32. In evaluating the effect of treatment on CD4, the best fitting marginal model was $E(Y_{ij}|A_i) = \beta_0 + \beta_1 A_i + \beta_2 t_{ij}$, where t_{ij} indicates the week of the j_{th} measurement on the i_{th} individual, and A_i was an indicator for the placebo arm. Since only follow-up measurements were modeled as outcomes and no interaction was detected between treatment and time, the effect of treatment was captured by β_1.

Standard, Simple Augmented, and Locally Efficient Augmented GEE were applied to estimate β_1. Several candidate outcome regression models for augmented GEE were identified through model selection procedures. Cross validation was used to select the final model, $E(Y_{ij}|A_i, X_i, t_i) = \eta_0 + \eta_1 A_i + \eta_2 t_{ij} + \eta_3 CD4_{0_i} + \eta_4 Sex_i$, where $CD4_0$ is baseline CD4. The QIC goodness-of-fit statistic {Pan and Wall (2002)} was compared among GEE fit to unaugmented marginal and conditional models to provide insight into the best fitting working covariance structures. To enforce compatibility of the marginal variance,

conditional variance, and outcome regression in fitting Locally Efficient Augmented GEE, the additive estimate of the marginal covariance was used. The working conditional variance was chosen by selecting the covariance structure resulting in the lowest QIC when fitting GEE on the conditional mean model. Simple augmented GEE were computed under all possible working marginal covariance structures, including the additive estimator motivated by the locally efficient GEE.

Table 2.4: QIC for selecting working covariance structures. Conditional model: $E(CD4_{ij}|Trt_i, Week_{ij}, \mathbf{X}_i) = \eta_0 + \eta_1 A_i + \eta_2 Week_{ij} + \eta_3 Sex_i + \eta_4 CD4_{0_i}$. Marginal model: $E(CD4_{ij}|Trt_i, Week_{ij}) = \eta_0 + \eta_1 Trt_i + \eta_2 Week_{ij}$

Conditional Model

Working Covariance Structure	QIC
Independence	1053.44
Exchangeable	1051.9
AR1	1052.29
Unstructured	1049.72

Marginal Model

Working Covariance Structure	QIC
Independence	1047.59
Exchangeable	1047.1
AR1	1046.56
Unstructured	1049.35

Regarding covariance selection, unstructured working covariance resulted in the lowest QIC for the conditional model (Table 2.4), suggesting the locally efficient estimator should be fit assuming an unstructured form of $V(\mathbf{Y}_i|\mathbf{X}_i, A_i)$. Results of the primary analysis are shown in Table 2.5. Several other covariance structures were also implemented for the locally efficient estimator to explore variance misspecification. Among simple augmented estimators, the additive marginal covariance resulted in lower variability than estimators using standard marginal covariance structures. Among Standard GEE with different working covariance models, the estimated difference in average CD4 for the placebo arm versus nelfinavir ranged from 9.9 to 20.17. The direction of the effect was reversed for estimators that incorporated baseline covariates, with average CD4 on the placebo arm 0.07 to 8.11 units lower than the nelfinavir arm. Treatment did not have a significant impact on CD4 at the 0.05 level for any of the estimators considered.

Estimators that incorporated baseline covariates greatly increased efficiency, with $SE(\hat{\beta}_1) \approx 20$ for Standard GEE and $SE(\hat{\beta}_1) \approx 9$ among augmented estimators (Relative Efficiency Augmented to Standard GEE $\approx$ 5.0). Simple augmented and locally efficient GEE resulted in similar efficiency. This may be due to the difficulty of correctly specifying all components of $\mathbf{h_{opt}}$, or because subjects had the same number of follow-up visits. As a benchmark for efficiency, we also fit unaugmented GEE assuming the conditional mean model $E(Y_{ij}|A_i, X_i, t_i) = \beta_0 + \beta_1 A_i + \beta_2 t_{ij} + \beta_3 CD4_{0_i} + \beta_4 Sex_i$ with an unstructured working covariance. This estimator represents the most efficient estimator of β_1 that may be obtained using $\mathbf{X}_i$, which requires assuming the more restrictive conditional mean model is correct. From this estimator, we can determine that for this particular case, there is little additional efficiency to be gained by locally efficient GEE if Simple Augmented GEE are fit under the best working covariance (Table 2.5)

Table 2.5: **Application of Standard, Simple Augmented, and Locally Efficient Augmented GEE to AIDS Clinical Trial Group Study 398.** Estimator (Working Marginal Covariance). Estimator: Unaugmented GEE (Standard), Simple Augmented GEE (Simple Aug. GEE), Locally Efficient Augmented GEE (Loc. Eff.). Working Marginal Covariance: Independence (Ind), Exchangeable (Exch), Autoregressive(1) (AR1), Unstructured (Un), Exchangeable for $V(E(Y|X,A)|A)$ and Unstructured for $V(Y|X,A)(Exch/Un)$, Exchangeable for $V(E(Y|X,A)|A)$ and AR1 for $V(Y|X,A)(Exch/AR1)$. Sandwich Standard Error (SE). Relative Efficiency (RE).

Estimator	$\hat{\beta}_1$	SE	RE
Standard (Ind)	9.971	20.772	0.942
Standard (Exch)	14.182	20.593	0.958
Standard (AR1)	16.977	20.222	0.993
Standard (Un)	20.173	20.156	1.000
Standard (Exch/Un)	14.615	20.347	0.981
Simple Aug. (Ind)	-8.110	9.203	4.797
Simple Aug. (Exch)	-6.385	8.904	5.124
Simple Aug. (AR1)	-3.059	9.244	4.754
Simple Aug. (Un)	-0.079	9.411	4.587
Simple Aug. (Exch/Un)	-5.972	8.571	5.530
Simple Aug. (Exch/AR1)	-5.048	8.920	5.106
Loc. Eff. (Ind)	-8.110	9.203	4.797
Loc. Eff. (Exch)	-6.821	8.953	5.068
Loc. Eff. (Exch/AR1)	-5.715	9.073	4.936
Loc. Eff. (Exch/Un)	-6.277	8.601	5.492
Adjusted (Un)	-6.649	8.621	5.467

2.5 Discussion

We derived and implemented a closed-form semiparametric locally efficient estimator of marginal treatment effects for correlated outcomes using baseline covariates. Through simulation, we demonstrated that our estimator is more efficient than corresponding suboptimal estimators in certain settings, particularly when randomized units vary in size and baseline covariates account for a large portion of the within-unit correlation. In longitudinal studies, variable size may occur when studies have staggered entry or as subjects are lost to follow-up. The estimator derived is only semiparametric locally efficient in the first case, as the locally efficient estimator for incomplete data incorporates information on the missingness process. More generally, large efficiency gains were shown for longitudinal analysis when the baseline level of the outcome was incorporated in estimation as an auxiliary covariate. Baseline levels of outcomes can be highly predictive of followup levels, suggesting that in the analysis of data from longitudinal studies, failing to incorporate baseline covariates in analysis can be extremely inefficient. Moreover, this motivates interest in developing methods for designing studies to incorporate baseline covariates in interim and final analyses.

Flexible Covariate-adjusted Exact Tests for Randomized Studies

Alisa J. Stephens, Eric J. Tchetgen Tchetgen, and Victor De Gruttola

Department of Biostatistics
Harvard University

3.1 Introduction

In randomized trials the primary goal is to evaluate the effect of a novel intervention on some outcome of interest. In addition to the treatment assignment and outcome, data on baseline covariates, such as demographics or biomarkers, are typically collected. To protect type I error, methods for including baseline covariates in analyses, whether as stratification factors or in regression models, are generally precisely defined. Recently, methods have been developed to allow for more flexible model selection without loss of protection of type error, at least asymptotically {Tsiatis et al. (2008); Zhang et al. (2008); Stephens et al. (2012a)}. Several studies have demonstrated that new methods permitting flexible use of baseline correlates of the outcome in analysis improve power and efficiency in treatment effect estimation {Tsiatis et al. (2008); Zhang et al. (2008); Stephens et al. (2012a)}. Nonetheless, in small samples, additional variability introduced by flexible model selection may fail to preserve type I error and also result in loss of power and efficiency compared to unadjusted analyses. In this paper, we evaluate several flexible covariate adjustment methods for studies with small numbers of randomized units. We examine the validity of adjusted tests through investigation of type I error and measure improvement over unadjusted tests by comparing power.

Consider a randomized trial in which n independent and identically distributed units $O_i = (Y_i, A_i, \mathbf{X}_i)$ are sampled from a population, where Y_i denotes the outcome of interest, A_i the random treatment assignment such that A_i=1, ..., K, and $\mathbf{X}_i$ the set of baseline covariates. For cluster-randomized or longitudinal trials, bold $\mathbf{Y}_i$ represents a multivariate outcome vector for individuals within the same randomized group or repeated measurements on a single randomized subject, respectively. In the context of multivariate outcomes, we consider settings where the treatment assignment is a scalar shared by measurements within the same cluster or subject. The primary analysis for most randomized trials compares outcomes Y_i among subjects assigned to different levels of treatment A_i. For scalar outcomes, tests comparing some feature of $f_{a^*}(Y)$, the distribution of Y under treatment a^*, are used to assess the statistical significance of observed differences in out-

comes across treatment groups. The two-sample t-test, Wilcoxon test, and their extensions for more than two groups are examples of commonly used methods. When outcomes are multivariate, modified versions of these tests are available to adjust standard errors for correlation among multiple measurements within the same randomized unit {Klar and Donner (2000)}.

Regression analysis may also be used to evaluate treatment effects. The effect of a binary treatment on the marginal mean of Y may be assessed through assuming the generalized linear model

$$E[Y_i|X_i, A_i] = g(\beta_0 + \beta_1 A_i), \tag{3.1}$$

where g^{-1} is a link function, and β is estimated through semiparametric estimating equations or fully parametric maximum likelihood inference. The effect of treatment on the marginal mean outcome $E[Y_i|A_i = a]$ is evaluated through testing $H_0 : \beta_1 = 0$. Under randomization, this effect is equivalent to no average causal effect of treatment on Y_i. When outcomes are multivariate, Y_i in (3.1) is replaced by Y_{ij}, denoting the j^{th} outcome of the i^{th} randomized unit for $i = 1, 2, ..., n$ and $j = 1, ...m_i$, where $M = \sum_{i=1}^{n} m_i$ is the total number of observations. For a semiparametric approach, generalized estimating equations (GEE) that account for correlation in responses may be used to obtain consistent parameter and standard error estimates. Regression methods naturally incorporate baseline covariates by assuming the adjusted mean model (AMM)

$$E[Y_i|\mathbf{X}_i, A_i] = g(\beta_0 + \beta_1^* A_i + \beta_X^{\mathrm{T}} \mathbf{X}_i). \tag{3.2}$$

When g is the identity link and the true model does not contain any treatment-covariate interactions, independence of A_i and $\mathbf{X}_i$ resulting from randomization guarantees that the adjusted estimator $\hat{\beta}_1^*$ is a consistent estimator of β_1. Moreover, it can be shown that $var(\hat{\beta}_1^*) \leq var(\hat{\beta}_1)$, where $\hat{\beta}_1$ is the unadjusted estimator, even under misspecification of the exact form of $\beta_X^{\mathrm{T}} \mathbf{X}_i$ in (3.2). For other link functions $\hat{\beta}_1^*$ is not consistent for β_1, nor does the addition of baseline covariates to the assumed mean model guar-

antee efficiency improvement. As a result, Zhang et al. (2008) advocate using a class of augmented estimators. Augmented estimators are derived from semiparametric theory and involve augmenting standard estimating functions by subtracting their Hilbert space projection onto the span of the scores of the treatment mechanism. Semiparametric theory provides theoretical justification for efficiency improvement of augmented estimators in large samples irrespective of the link function g and only assuming model (3.1) holds. Stephens et al. (2012a) demonstrated how augmentation may be used for clustered or longitudinal data by augmenting generalized estimating equations. The same authors also presented the locally efficient augmented estimator Stephens et al. (2012b) under model (3.1). Augmented inference relies on asymptotic theory and therefore requires a fairly large number of randomized units. In large samples, model selection variability for baseline covariates is small provided the number of covariates is not large; in small samples, however, flexible covariate selection induces additional variability that may lead to variance underestimation and loss of efficiency.

In contrast, Rosenbaum (2002) extended the randomization theory of Fisher (1935) to propose an exact covariate-adjusted test that does not assume a particular distribution for outcomes or that the observed data are a random sample from some unobserved population of independent units. Randomization inference considers independent units $\tilde{O}_i = (y_i, A_i, \mathbf{x}_i)$, where the lowercase notation emphasizes that only the treatment assignment A_i is random, and outcomes y_i and baseline covariates $\mathbf{x}_i$ are fixed. The exact method tests the sharp null hypothesis $H_0 : y_a = y_*$ for all a, where y_a is a subject's potential outcome under treatment a. Rosenbaum (2002) discussed the potential outcomes framework in detail. The null distribution of the test statistic is obtained through permutation of A_i. The test proposed by Gail et al. (1988) approximates the exact test by standardizing the observed test statistic by its randomization-based variance and comparing to the standard normal distribution. Post model-selection inference based on the Gail et al. and Rosenbaum approaches have not been investigated; we consider settings where model selection is used to determine covariates that explain variability in y_i. Adaptive selection of baseline covariates may be particularly useful when $\mathbf{x}_i$ is high-dimensional or

prior knowledge is not available to inform covariate adjustment. Further improvement in small-sample inference may be possible from higher order approximations of the distribution of a class of randomization test statistics {Bickel and Zwet (1978)}, but this theory has not yet been evaluated in practice.

Details of the four covariate-adjusted tests: I) Adjusted mean models (AMM), II) Augmented marginal model, III) Approximate exact, and IV) Exact (permutation) are discussed in Section 3.2. Inference for independent and correlated outcomes is presented. In Section 3.3, the small sample properties of covariate-adjusted tests are evaluated through simulation. Methods are illustrated through application to the *Young Citizens* study in Section 3.4. Finally, we summarize our results and provide recommendations for practical use in Section 3.5.

3.2 Methods

We consider four methods of covariate-adjusted hypothesis testing: I) Wald test of β_1^* in the adjusted mean model (3.2), II) Wald test of β_1 in marginal model (3.1), in which estimating equations are augmented to include baseline covariates, III) approximate exact test, and IV) the exact test. The presentation is not exhaustive for covariate-adjusted inference but considers widely recognized classical and modern methods. Each test is first presented for independent outcomes and followed by generalizations for dependent data.

3.2.1 Independent Outcomes

Method Ia: Wald test of β_1^* in model (3.2)

Assuming model (3.2) holds, parameters β and respective standard errors are estimated via maximum likelihood or semiparametric estimating equations. The null hypothesis $H_0 : \beta_1^* = 0$ is evaluated through the test statistic $T_c = \frac{\hat{\beta_1}^*}{SE(\hat{\beta_1}^*)}$.

Method IIa: Wald test of β_1 in model (3.1) with augmented estimating equations {Tsiatis et al. (2008),Zhang et al. (2008)}

Unlike inference on the AMM, the augmentation method assumes model (3.1). Predicted values from a working model for the conditional mean $E[Y_i|\mathbf{X}_i, A_i]$ are incorporated in estimating equations that are solved to estimate β. Consistent estimates of β_1 are obtained even if $E[Y_i|\mathbf{X}, A_i]$ is misspecified.

To test $H_0 : \beta_1 = 0$, the test statistic $T_a = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$ is considered, where $\hat{\beta}_1$ is the solution of the augmented estimating equations

$$\sum_{i=1}^{n} \psi_a(O_i;\beta) = \sum_{i=1}^{n} \Bigg[h(A_i;\beta)\{Y_i - g(A_i;\beta)\} - \sum_{a=1}^{K} \{I(A_i = a) - \pi_a\}\{h(a;\beta)(E[Y_i|X_i, A_i = a] - g(a;\beta))\} \Bigg] = \mathbf{0}, \quad (3.3)$$

where π_a denotes $P(A_i = a)$. In practice ψ_a is evaluated by $\hat{\psi}_a$, where the true regression $E[Y_i|\mathbf{X}_i, A_i = a]$ is approximated by the working model $E[Y_i|\mathbf{X}_i, A_i = a] = d(\mathbf{X}_i; \eta_a)$ evaluated under an estimate $\hat{\eta}_a$. The subscript a emphasizes that the regression for augmented estimators conditions on the treatment assignment. Alternatively, $E[Y_i|\mathbf{X}_i, A = a]$ may be estimated separately in each treatment arm, resulting in K regression models that do not contain indicators for treatment. The variance of $\hat{\beta}_1$ is estimated by the sandwich variance estimator $\hat{Var}(\hat{\beta}_1) = C \left[\left(\sum_{i=1}^{n} \frac{dh(A_i;\beta)}{d\beta^{\mathrm{T}}} D_i \right)^{-1} \sum_{i=1}^{n} \left[\psi_a(O_i;\beta)^{\otimes 2} \right] \left(\sum_{i=1}^{n} \frac{dh(A_i;\beta)}{d\beta^{\mathrm{T}}} D_i \right)^{-1} \right]$, where $D_i = \frac{dg(A_i;\beta)}{d\beta^{\mathrm{T}}}$, and $C = \{(n_0 - p_0 - 1)^{-1} + (n_1 - p_1 - 1)^{-1}\}/\{(n_0 - 1)^{-1} + (n_1 - 1)^{-1}\}$ is incorporated to account for finite-sample variability attributable to augmenting. In C, n_a is the sample size in treatment arm a and p_a is the dimension of η_a for the working covariate-adjustment model.

Method IIIa: Approximation of the Exact Test

The approximation of the exact test considers the $H_0 : y_a = y_*$ for all a. To test H_0

we construct the test statistic

$$T_s = \frac{S}{\sqrt{Var(S|y,\mathbf{x})}}, \text{ where } S = \sum_{i=1}^{n}(A_i - \pi)w_i \tag{3.4}$$

and $Var(S|y,\mathbf{x})$ is shown in (3.5). Baseline covariates are incorporated by setting $w_i = \hat{\varepsilon}_i = y_i - d(\mathbf{x}_i;\hat{\eta})$, the residual from the working mean model $d(\mathbf{x}_i;\hat{\eta})$, which estimates the true regression model $E[y_i|\mathbf{x}_i] = f(\mathbf{x}_i;\eta)$ under the sharp null. For unadjusted analysis, w_i=y_i. We purposely omit the subscript a on the regression function as a reminder that under the sharp null, y_i cannot depend on treatment, so A_i is excluded from the proposed working model. The variance of S is calculated by

$$Var(S|y,\mathbf{x}) = \pi(1-\pi)\sum_{i=1}^{n} w_i^2 + \overbrace{\left(\pi\frac{n/2-1}{n-1} - \pi^2\right)\sum_{i\neq i'} w_i w_{i'}}^{(Q)}, \tag{3.5}$$

and significance is determined by comparing $|T_a|$ to the standard normal distribution.

Term Q in $Var(S|\mathbf{y},\mathbf{x})$ is nonzero when the total number of subjects assigned to each treatment is fixed. This typically applies in trials with small samples, where matching and blocked randomization strategies are employed to prevent imbalances in treatment allocation that may occur with unrestricted random assignment. Under such randomization, the vector $\mathbf{A} = (A_1, A_2, ..., A_n)$ follows a hypergeometric distribution, where the probability of being assigned to treatment for a particular subject is affected by the other subjects' treatment assignments. When w_i is the residual from a working model for $E[y_i|\mathbf{x}_i]$, Q ≈ 0, as $E[\varepsilon_i|\mathbf{x}_i]$=0, and $\varepsilon_i \perp \varepsilon_{i'}$. If considering the unadjusted outcomes Y_i, failure to include Q may result in gross variance overestimation and extremely conservative testing for small n. In large samples, Q ≈ 0 for $w_i = \hat{\varepsilon}_i$ or $w_i = y_i$.

For the class of statistics defined by $T = \sum_{i=1}^{n} A_i c_i$, where c_i is a score, Bickel and Zwet (1978) determined a higher-order approximation for the distribution of the standardized statistic T^*, given by

$$P(T^* < t) = \Phi(t) - \frac{\phi(t)}{\pi(1-\pi)}\left[\frac{\pi(1-\pi)}{2n}H_1(t) + \frac{\sqrt{\pi(1-\pi)}(1-2\pi)}{6}\frac{\sum_{i=1}^{n}(c-c_.)^3}{\left\{\sum_{i=1}^{n}(c-c_.)^2\right\}^{3/2}}H_2(t)+\right.$$
$$\left.\left\{\frac{1-6\pi+6\pi^2}{24}\frac{\sum_{i=1}^{n}(c-c_.)^4}{\left\{\sum_{i=1}^{n}(c-c_.)^2\right\}^{2}} - \frac{(1-2\pi)^2}{8n}\right\}H_3(t) + \frac{(1-2\pi)^2}{72}\frac{\left\{\sum_{i=1}^{n}(c-c_.)^3\right\}^{2}}{\left\{\sum_{i=1}^{n}(c-c_.)^2\right\}^{3}}H_5(t)\right] \tag{3.6}$$

The expansion suggests that a higher order accurate quantile of the distribution of the test statistic may be found by solving for Z^*_α such that $P(T < Z^*_\alpha) = 1 - \alpha/2$ for two-sided tests.

Method IVa: Exact Test

The exact test also applies to the hypothesis $H_0 : y_a = y_*$ for all a; the null distribution of $T_p = S$ is calculated by permuting the treatment assignment A_i among subjects. For each permutation, the test statistic T_p is calculated under the permuted treatment assignment A_b, resulting in distribution of statistics $T_p(A_b)$. The exact null distribution is often estimated by considering B permutations for large B, and a p-value is obtained by $p_B = \frac{1}{B} = \sum_{b=1}^{B} I(|T_p(A_b)| > |T_p|)$. For a level α test, we reject the sharp null of no treatment effect when $p_B < \alpha$.

3.2.2 Dependent Outcomes

For clustered outcomes, we consider modifications of the univariate tests that accommodate correlation in responses.

Method Ib: Wald test of β^*_1 in model (3.2) using GEE {Liang and Zeger (1986)}

To accommodate correlation in outcomes within a cluster, generalized estimating equations may be constructed assuming model (3.2) holds. The adjusted treatment effect β_1^* is estimated by solving the generalized estimating equations

$$\sum_{i=1}^{n} \mathbf{D}_i \mathbf{V}_i^{-1} [\mathbf{Y}_i - \mathbf{g}(A_i, \mathbf{X}; \beta)] = \mathbf{0}, \tag{3.7}$$

where $\mathbf{D}_i = \frac{d\mathbf{g}(A_i, \mathbf{X}; \beta))}{d\beta^{\mathrm{T}}}$, $\mathbf{V}_i = V_i(\phi)^{1/2} \mathbf{R} V_i(\phi)^{1/2}$. The working covariance $\mathbf{V}_i$ is determined by the $m_i \times m_i$ correlation matrix $\mathbf{R}$ and diagonal variance matrix $V_i(\phi)$. The variance of $\hat{\beta}$ is calculated by the sandwich variance estimator,

$$\hat{var}(\hat{\beta}) = \left(\sum_{i=1}^{n} \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \left(\sum_{i=1}^{n} \left[\mathbf{D}_i \mathbf{V}_i^{-1} \{\mathbf{Y} - \mathbf{g}(A_i, \mathbf{X}; \beta)\} \right]^{\otimes 2} \right) \left(\sum_{i=1}^{n} \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1}, \tag{3.8}$$

and T_c is calculated to evaluate $H_0 : E[\mathbf{Y}_i | \mathbf{X}_i, A_i = 1] = E[\mathbf{Y}_i | \mathbf{X}_i, A_i = 0]$.

Method IIb: Wald test of β_1 in model (3.1) using augmented GEE {Stephens et al. (2012a)}

Assuming marginal model (3.1), augmented estimating equations are formed by

$$\sum_{i=1}^{n} \psi_a(O_i; \beta, \eta) = \sum_{i=1}^{n} \left\{ \mathbf{D}_i \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \mathbf{g}(A_i; \beta)\} - \sum_{a=1}^{K} \{I(A_i = a) - \pi_a\} [\mathbf{D}_i(a) \mathbf{V}_i^{-1}(a) \{\mathbf{d}\{\mathbf{X}_i; \eta_a) - \mathbf{g}(a; \beta)\}] \right\} = \mathbf{0}, \tag{3.9}$$

where $\mathbf{d}(\mathbf{X}_i; \eta_a)$ is an estimate of $E[\mathbf{Y}_i | A_i = a, \mathbf{X}_i]$. To estimate $var(\hat{\beta})$, the standard estimating function is replaced with the augmented estimating function ψ_a in the middle term of (3.8).

Method IIIb: Approximation to the Exact Test (Multivariate)

Although responses y_{ij} and covariates $\mathbf{x}_{ij}$ are considered fixed for randomization inference, the calculated covariance among y_{ij} in the i^{th} cluster incorporates information

on the difference in the between versus within sum of squares, which may increase power in testing. A working covariance $\mathbf{V}_i$ as for GEE is incorporated into testing by

$$S_D = \sum_{i=1}^{n} (A_i - \pi)\mathbf{1}\mathbf{V}_i^{-1}\mathbf{w}_i, \tag{3.10}$$

where $\mathbf{w}_i$ is the residual vector $\mathbf{w}_i = (w_{i1}, w_{i2}, ..., w_{im_i})^{\mathrm{T}}$ determined by $w_{ij} = \hat{\varepsilon}_{ij} = y_{ij} - d(\mathbf{x}_{ij}; \hat{\eta})$ and $\mathbf{1}$ is the m_i-dimensional vector of 1s. To estimate correlation parameters, the method of moments is used. We consider the moment estimating equations

$$\sum_{i=1}^{n} \sum_{j<j'} \left\{ \frac{w_{ij} w_{ij'}}{\tau} - r(\gamma) \right\}, \tag{3.11}$$

where $\tau = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}^2$. The weight matrix $\mathbf{V}_i$ is given by $V_i = L^{1/2} U L^{1/2}$, where L is an $m_i \times m_i$ diagonal matrix with τ along the diagonal, and $\mathbf{U}$ is a correlation matrix, where $Q_{j,j'} = r(\gamma)$. For vector-valued outcomes $\mathbf{Y}_i$, the variance is

$$Var(S|\mathbf{y_i}, \mathbf{x_i}) = \pi(1-\pi) \sum_{i=1}^{n} (\mathbf{1}\mathbf{V}_i^{-1}\mathbf{w}_i)^{\otimes 2} + \left(\pi \frac{n/2 - 1}{n-1} - \pi^2 \right) \overbrace{\sum_{i \neq i'} (\mathbf{1}\mathbf{V}_i^{-1}\mathbf{w}_i)(\mathbf{1}\mathbf{V}_{i'}^{-1}\mathbf{w}_{i'})^{\mathrm{T}}}^{Q*}, \tag{3.12}$$

where $Q*$ is the small sample correction for fixed treatment allocation. Bickel and Zwet (1978) may be applied to dependent outcomes as well to ensure nominal type I error levels in small samples.

Method IVb: Exact Test (Multivariate)

The null distribution of test statistic (3.10) is determined by permuting the cluster-level treatment assignment A_i. Because outcomes and covariates are fixed, the residuals $\hat{\varepsilon}_{ij} = y_{ij} - d(\mathbf{x}_{ij}; \hat{\eta})$ and working covariance $\mathbf{V}_i$ do not depend on the permuted treatment assignment under H_0. Working covariance parameters therefore only need to be estimated once, and $\mathbf{V}_i$ is equal for all permutations A_b. Significance is established as in section 3.2.1.

3.2.3 Model Selection for Baseline Covariates

When the set of baseline covariates is high dimensional, adjusting for all available covariates may be inefficient. Prior knowledge may suggest the inclusion of some covariates; among other covariates whose impact on Y_i is not well understood model selection may help to determine which covariates to include. Adjusted mean models and augmented estimation require the conditional mean model $E[Y|\mathbf{X}, A]$, whereas randomization inference requires an estimate of $E[Y|\mathbf{X}]$. Current literature provides a wide array of methods for selection of baseline covariates, particularly for univariate outcomes. Stepwise selection procedures based on some entry criterion may be used. Methods based on penalized likelihoods such as LASSO {Tibshirani (1996)}, adaptive LASSO {Zou (2006)}, SCAD {Fan and Li (2001)}, and MC+ {Zhang (2010)} are equivalently applicable. Model selection for multivariate outcomes is less developed, but extensions of available methods are presented and discussed in Sofer et al. (2012). We consider two popular approaches, forward selection by AIC or BIC, and adaptive LASSO, {Zou (2006)} where the tuning parameter is selected by cross validation.

Forward selection is an example of a greedy algorithm, defined as an algorithm that makes the locally optimal choice at each stage in search of a global optimum {Black (2005)}. To find the best predictive model, forward selection starts with a generalized linear model containing the intercept and at each step enters a single covariate according to a prespecified criterion. Examples of entry criteria include minimizing p-values or an information criterion such as AIC, or maximizing adjusted r^2.

Model selection by penalized regression is derived by minimizing an objective function

$$\Omega(\beta) = \sum_{i=1}^{n} L\{Y_i, g(A_i, \mathbf{X}; \beta)\} + P_\lambda(\beta), \tag{3.13}$$

which consists of a loss function $L\{Y_i, g(A_i, \mathbf{X}; \beta)\}$ and a penalty $P_\lambda(\beta)$, where $P_\lambda(\beta)$ is indexed by a nonnegative tuning parameter λ. The form of $P_\lambda(\beta)$ defines various regularized regression methods; for adaptive LASSO $P_\lambda(\beta) = \lambda \sum_{k=1}^{p} \hat{w}_k |\beta_k|$ with weights

$\hat{w}_k = 1/|\hat{\beta}_k^{\gamma}|$ derived from an initial fit of β. We consider an adaptive LASSO-hybrid implementation motivated by the LASSO-OLS hybrid {Efron et al. (2004)}, in which LASSO is used to determine the covariates for which $\beta_k \neq 0$, and the selected model is subsequently fit by OLS.

When outcomes are multivariate Sofer et al. (2012) discusses that accounting for correlation improves the efficiency of penalized regression estimates. In small samples, it is especially desirable to reduce the variability in penalized regression since the number of units may not be sufficient to achieve consistency despite estimation under a misspecified independence correlation structure. The authors recommend scaling outcomes and covariates by $\Lambda^{1/2}$, where Λ=$\mathbf{V}_i^{-1}$ is a working precision matrix based on an initial estimate of the coefficient vector. The initial estimate may be determined by a model selection method that assumes independence. For validation-based penalized regression, estimation proceeds as in the univariate case on the scaled outcomes $\tilde{\mathbf{Y}}_i = \Lambda^{1/2}\mathbf{Y}_i$ and covariates $\tilde{\mathbf{X}}_i = \Lambda^{1/2}\mathbf{X}_i$. We also consider forward selection of $\tilde{\mathbf{Y}}_i$ on $\tilde{\mathbf{X}}_i$ to evaluate possible improvements in model selection and resulting power for testing treatment effects.

3.3 Simulation Study

3.3.1 Univariate

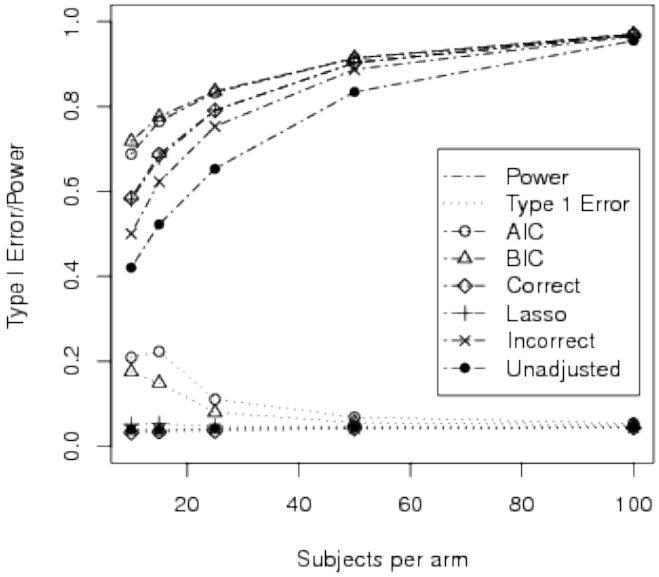
We first consider scalar outcomes Y_i. For each simulated dataset 25 baseline covariates $X_{i_1}, ..., X_{i_{25}}$ were generated from the multivariate lognormal distribution by exponentiating draws from the multivariate normal distribution with mean $\mu = (0, 0, ..., 0)$ and covariance Σ, where Σ was defined such that $corr(\log(X_{i_k}),\log(X_{i_{k'}})) = 0.5$ for $k, k' = 1, ..., 10$, $corr(\log(X_{i_k}),\log(X_{i_{k'}})) = 0.2$ for $k = 1, ..., 10, k' = 11, ..., 20$, $corr(\log(X_{i_k}),\log(X_{i_{k'}}) = 0$ for $k = 1, ..., 20, k' = 2, ..., 25$, and $var(log(X_{i_k})) = 1$ for $k = 1, ..., 25$. Treatment A_i was binary and simulated with a fixed, equal number of subjects assigned to treatment or control. Outcomes were generated from the model $Y_i = \eta_0 + \eta_1 A_i + \eta_2 X_{i_1} + \eta_3 X_{i_2} + \eta_4 X_{i_{10}} + \eta_5 X_{i_{11}} \eta_6 X_{i_{12}} + \varepsilon_i$ with $\log(\varepsilon_i) \sim N(0, 1.9)$,

$\eta' = (1, 0, 1, 1, 0.2, 0.2, 0.2)$ under the null and $\eta' = (1, 4, 1, 1, 0.2, 0.2, 0.2)$ under the alternative. Sample sizes of $n_a = 10, 15, 25, 50, 100$ in each treatment arm were considered. Under this design, baseline covariates accounted for roughly 30% of the variability in $Y_i|A_i$.

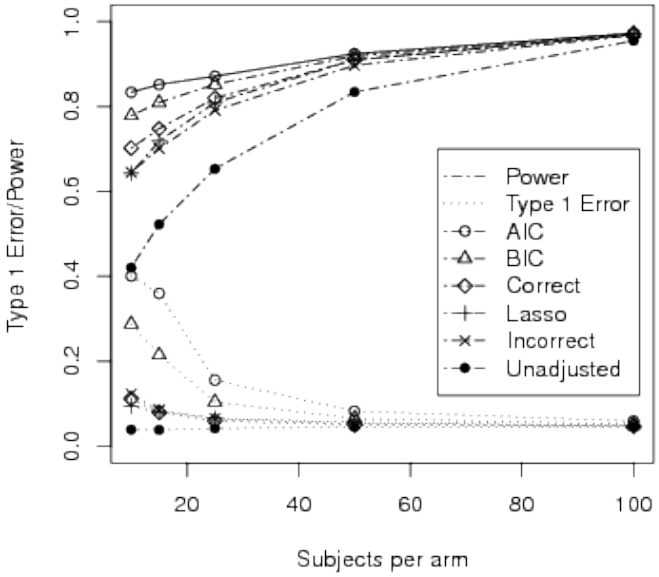
All four covariate-adjusted methods were applied to each dataset, and various adaptive procedures were used to select among the 25 baseline covariates. Several variations for each covariate-adjusted test were considered, with each variation defined by a different regression model. For adaptive approaches, selection of regression models was based on three different methods: forward selection minimizing AIC, forward selection minimizing BIC, and the adaptive LASSO-OLS hybrid. The adaptive LASSO tuning parameter was selected by l-fold cross validation, where $l = n/10$. For Method Ia, inference was performed by OLS on the model including A_i and covariates suggested by the adaptive model selection procedure. Adaptively selected models were compared to two fixed models: the data generating model, which serves as a benchmark for the largest possible improvement in power, and an incorrect model, $E[Y_i|\mathbf{X}_i, A_i] = \eta_0 + \eta_1 X_{i_1} + \eta_2 X_{i_3} + \eta_3 X_{10} + \eta_4 X_{i_{13}} + \eta_5 X_{i_{21}}$, including two predictive covariates and 3 noisy covariates. Finally, each method was also applied to the unadjusted outcomes Y_i to assess whether incorporating baseline covariates improved power compared to no adjustment. Treatment was forced into the regression model for Methods Ia and IIa. Considering Methods IIIa and IVa, treatment was omitted from covariate selection, as the sharp null excludes any estimated effect of treatment, even if not significant. In addition to assessing type I error and power when the true data-generating model was contained in the set of candidate models, we assessed power when important transformations for baseline covariates were not included. We modified the data generating mechanism to include squared terms for X_{i_1} and $X_{i_{10}}$ and changed the coefficient of X_{i_1} to $\eta_1 = 0.50$. As in the previous setting, model fitting algorithms for determining predictive covariates only considered linear terms.

Figure 3.1: **Type I Error and Power of Univariate AMM and Augmented Tests.** Adaptive regression model selection: AIC, BIC, Adaptive LASSO. Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Figure 3.1(Continued)



(a) AMM



(b) Augmented

Table 3.1: **Type I Error of Univariate Covariate-adjusted Tests.** Adjusted mean model (AMM), Augmented, Approx. Exact (without Bickel adjustment), Approx. Exact (Sm) (with Bickel adjustment) and Exact tests. Adaptive regression model selection: AIC, BIC, Adaptive LASSO (A. LASSO). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Adjusted Mean Model

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.0384	0.2089	0.1744	0.0505	0.0381	0.0311
15	0.0379	0.2224	0.1488	0.0526	0.037	0.0333
25	0.0414	0.1102	0.0792	0.0465	0.04	0.0344
50	0.0444	0.0679	0.055	0.0464	0.0407	0.0409
100	0.0445	0.053	0.0486	0.044	0.043	0.0425

Augmented

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.0384	0.4005	0.2874	0.0936	0.1228	0.1116
15	0.0379	0.3595	0.2143	0.0801	0.0846	0.0788
25	0.0414	0.1551	0.1036	0.0652	0.0645	0.0588
50	0.0444	0.082	0.0649	0.0559	0.0524	0.0493
100	0.0445	0.0585	0.051	0.0462	0.0486	0.0466

Approx. Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.0346	0.0368	0.0383	0.0356	0.0368	0.0356
15	0.0354	0.0446	0.0406	0.0375	0.0347	0.0375
25	0.0398	0.0375	0.0388	0.039	0.0389	0.039
50	0.0438	0.0415	0.0423	0.0417	0.0398	0.0417
100	0.0442	0.0421	0.0438	0.0418	0.043	0.0418

Approx. Exact (Sm)

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.033	0.035	0.036	0.034	0.0347	0.034
15	0.0354	0.0442	0.0396	0.037	0.0345	0.037
25	0.0412	0.0384	0.0398	0.0403	0.0394	0.0403
50	0.0456	0.0433	0.0443	0.0432	0.0424	0.0432
100	0.0454	0.0432	0.0453	0.0433	0.0442	0.0433

Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.0498	0.0487	0.0491	0.0489	0.0519	0.0486
15	0.0499	0.0543	0.0511	0.0491	0.0481	0.0495
25	0.0518	0.0456	0.0491	0.0492	0.0509	0.0494
50	0.0515	0.0517	0.0529	0.0541	0.0524	0.0546
100	0.0505	0.0483	0.0524	0.0489	0.0513	0.0504

Table 3.2: **Power of Univariate Covariate-adjusted Tests when the correct model is a candidate model.** Adjusted mean model (AMM), Augmented, Approx. Exact (without Bickel adjustment), Approx. Exact (Sm) (with Bickel adjustment) and Exact tests. Adaptive regression model selection: AIC, BIC, Adaptive LASSO (A. LASSO). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Adjusted Mean Model

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.4204	0.6883	0.7182	0.5805	0.4999	0.5843
15	0.5224	0.7647	0.7758	0.6796	0.6226	0.6871
25	0.6532	0.8329	0.8362	0.791	0.7532	0.7912
50	0.8343	0.9139	0.9144	0.9035	0.8874	0.9029
100	0.9549	0.9706	0.971	0.9692	0.9658	0.9687

Augmented

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.4204	0.7991	0.7786	0.643	0.6448	0.7018
15	0.5224	0.8244	0.8095	0.7188	0.7012	0.7476
25	0.6532	0.8573	0.8523	0.8091	0.7911	0.82
50	0.8343	0.9206	0.9188	0.9102	0.8971	0.9096
100	0.9549	0.9722	0.9722	0.97	0.9679	0.9705

Approx. Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.4091	0.365	0.4649	0.5136	0.4761	0.5586
15	0.515	0.4567	0.6038	0.6316	0.6116	0.6793
25	0.6494	0.7351	0.7819	0.7718	0.7486	0.7891
50	0.8339	0.8957	0.9034	0.8983	0.8868	0.9029
100	0.9547	0.9683	0.9686	0.9682	0.9657	0.9686

Approx. Exact (Sm)

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.4051	0.3567	0.4552	0.5056	0.4681	0.5549
15	0.5139	0.4528	0.5996	0.6297	0.6107	0.6807
25	0.6516	0.7366	0.7831	0.7741	0.7515	0.7922
50	0.8358	0.898	0.9055	0.9014	0.8901	0.9054
100	0.9562	0.9695	0.971	0.9696	0.9676	0.97

Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.4594	0.393	0.4951	0.5486	0.5104	0.5934
15	0.5529	0.4753	0.6198	0.6594	0.6409	0.7074
25	0.6781	0.752	0.7973	0.7955	0.7734	0.8151
50	0.8465	0.9059	0.914	0.9126	0.8998	0.9171
100	0.9618	0.9747	0.9752	0.9752	0.9734	0.9759

Results for type I error are shown in Figure 3.1 and Table 3.1. Method Ia performed poorly for small sample sizes with model selection, leading to type I error rates as large as α=0.2. For fixed models chosen apriori, testing β_1^* preserves type I error, and is even slightly conservative as a result of the skewness in the covariates and outcomes (α=0.0311-0.043). The performance of asymptotically equivalent Method IIa varies over the choice of model selection procedure. For adaptive LASSO, the augmented test resulted in type I errors roughly twice the nominal level at $n_a = 10$. Adaptive selection of covariates by AIC or BIC had even larger type I error inflation (α=0.40 for n_a=10). Type I error was still not preserved when augmenting with fixed models (0.12 for n_a=10). By contrast, Methods IIIa and IVa maintained type I error at all sample sizes considered. The approximate exact test remained slightly conservative due to skewness, while the exact test achieved nominal type I error levels. There are noteworthy differences in the behavior of the various model selection procedures. As expected, BIC favored more parsimonious models than AIC: AIC-based selection resulted in models with 5 to 7 baseline covariates on average; BIC, 3 to 4 covariates. Adaptive LASSO was the most conservative model selection procedure, including 1 to 4 covariates on average, with the number of covariates selected increasing with the sample size.

Table 3.2 provides simulation results demonstrating the impact of model selection procedures on power. For $n_a \leq 50$, covariate adjustment based on AIC and BIC resulted in larger power than did the correct covariate adjustment model for Methods Ia and IIa (Power=0.68-0.91 for AIC and BIC, Power=0.58-0.90 for the correct model), suggesting that the former led to overfitting of the regression. The power of adjustment with adaptive LASSO did not exceed the power of adjustment under the correct model for any covariate-adjusted test statistic considered. In general, Methods IIIa and IVa had lower power than Methods Ia and IIa, reflecting the fact that the randomization-based tests preserve type I error while adding covariates to the mean model and augmentation do not. For very small sample sizes ($n_a \leq 15$), covariate adjustment by AIC lost power relative to the unadjusted test (Approx. Exact AIC = 0.36-0.46 , Approx. Exact Unadjusted 0.41-0.52 ; Exact AIC = 0.49-0.57, Exact Unadjusted = 0.59-0.64). For $n_a \geq 25$, AIC-based adjust-

Table 3.3: **Average Number of Baseline Covariates** selected by AIC, BIC, and Adaptive LASSO by sample size when candidate models include the correct model. First entry - number of baseline covariates selected when treatment was forced into the model. Second entry - number of baseline covariates when treatment was omitted from the model.

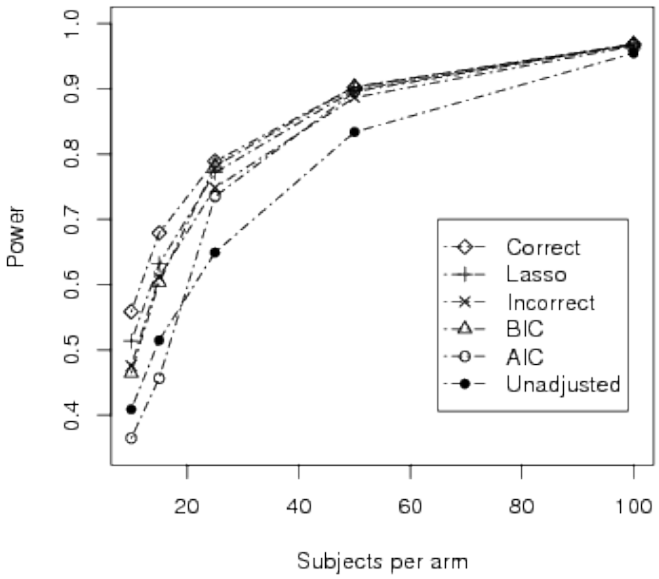
n_a	AIC	BIC	A. LASSO
10	6.45	3.93	1.84
	5.75	3.60	1.61
15	8.65	4.14	2.63
	7.96	3.93	2.26
25	6.13	3.19	3.11
	5.95	3.15	2.87
50	5.46	2.94	3.69
	5.41	2.93	3.57
100	5.49	3.01	3.92
	5.48	3.00	3.82

ment improved power compared to no adjustment. Model selection by BIC and adaptive LASSO, which penalize more severely for model complexity than AIC, improved power over unadjusted test statistics across all simulated sample sizes. Method IVa had higher power than Method IIIa, with the difference in power increasing inversely with sample size. Across all settings considered, Bickel's adjustment for the distribution of the approximate exact test had little impact on resulting inferences, suggesting even higher order terms may be necessary to recover nominal type I error.

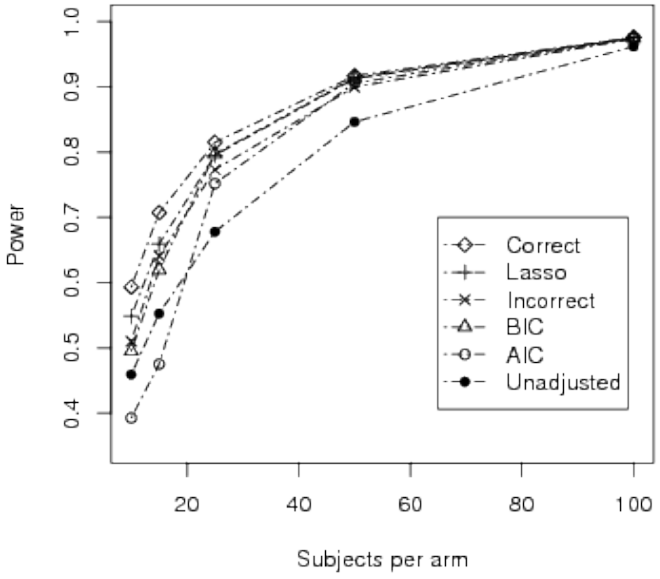
In the second set of power simulations, the data-generating model contained quadratic terms that were not considered in covariate adjustment. Results are shown in Figure 3.3-3.3 and Table 3.4. The relative performance of adaptive procedures remained the same. At small samples sizes, exact inference AIC resulted in less power improvement than the other adjustment methods. At $n_a = 10$, exact inference based on the AIC-selected model mirrored unadjusted exact inference (Method IVa AIC = 0.27, Method IVa Unadjusted=0.25). Considering Method IIIA, AIC-based inference increased power relative to not adjusting, but gains were limited compared to BIC selection, adaptive LASSO, and the prespecified incorrect model (AIC =0.245, Unadjusted= 0.18, BIC=0.3166, adaptive LASSO=0.3541, Prespecified=0.3044). Increasing the sample size per arm to $n_a = 25$, power for AIC-selected adjustment was more similar to the BIC and adaptive LASSO. At

Figure 3.2: **Power of Univariate Approx. Exact and Exact Tests when the correct model is a candidate model.** Adaptive regression model selection: AIC, BIC, Adaptive LASSO. Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Figure 3.2(Continued)



(a) Approximate

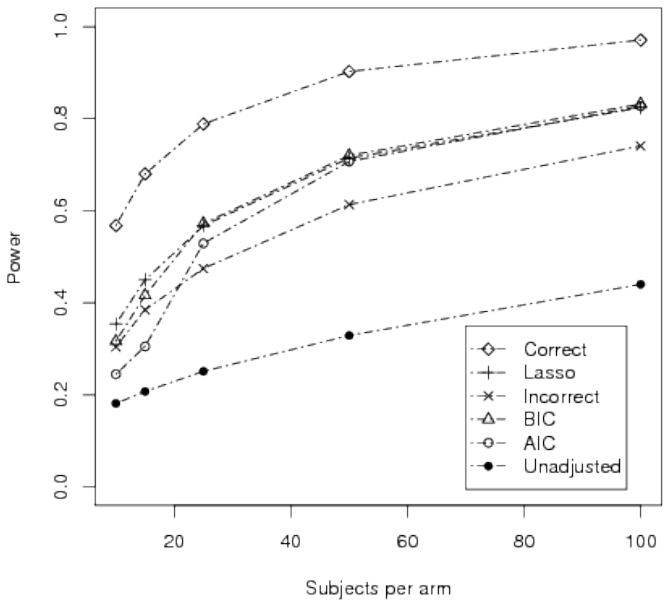


(b) Exact

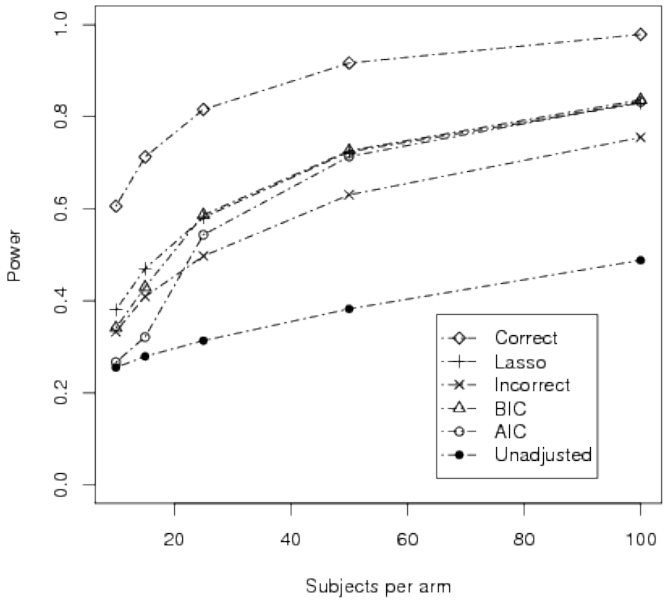
$n_a \geq 50$, all adaptive procedures resulted in similar power, while the incorrect prespecified model had lower power (Prespecified=0.49-0.75, Adaptive Methods = 0.54-0.84).

Figure 3.3: **Power of Univariate Approx. Exact and Exact Tests when the correct model is not a candidate model.** Adaptive model selection: AIC, BIC, Adaptive LASSO. Pre-specified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Figure 3.3(Continued)



(a) Approximate



(b) Exact

Table 3.4: **Power of Univariate Covariate-adjusted Tests when the correct model is not a candidate model.** Adjusted mean model (AMM), Augmented, Approx. Exact (without Bickel adjustment), Approx. Exact (Sm) (with Bickel adjustment) and Exact tests. Adaptive regression model selection: AIC, BIC, Adaptive LASSO (A. LASSO). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Adjusted Mean Model

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.1880	0.5805	0.6094	0.4500	0.5883	0.3231
15	0.2132	0.6545	0.6557	0.5359	0.6894	0.3947
25	0.2544	0.6809	0.6692	0.6150	0.7919	0.4793
50	0.3305	0.7613	0.7554	0.7343	0.9030	0.6154
100	0.4413	0.8417	0.8412	0.8295	0.9714	0.7419

Augmented

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.1880	0.7467	0.7057	0.5427	0.6054	0.4729
15	0.2132	0.7556	0.7143	0.6067	0.6397	0.4889
25	0.2544	0.7297	0.7078	0.6588	0.7134	0.5329
50	0.3305	0.7820	0.7701	0.7508	0.8196	0.6386
100	0.4413	0.8480	0.8476	0.8367	0.9071	0.7512

Approx. Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.1815	0.2450	0.3166	0.3541	0.5680	0.3044
15	0.2075	0.3053	0.4161	0.4501	0.6800	0.3847
25	0.2512	0.5292	0.5724	0.5673	0.7884	0.4746
50	0.3290	0.7069	0.7204	0.7142	0.9025	0.6133
100	0.4401	0.8269	0.8322	0.8238	0.9710	0.7409

Approx. Exact (Sm)

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.1792	0.2380	0.3101	0.3452	0.5629	0.2995
15	0.2088	0.3020	0.4114	0.4479	0.6820	0.3829
25	0.2569	0.5284	0.5719	0.5676	0.7915	0.4760
50	0.3360	0.7075	0.7219	0.7153	0.9059	0.6166
100	0.4499	0.8289	0.8328	0.8250	0.9726	0.7442

Exact

n_a	Unadjusted	AIC	BIC	A. LASSO	Incorrect	Correct
10	0.2669	0.3412	0.3803	0.6056	0.3329	0.2551
15	0.3212	0.4298	0.4700	0.7127	0.4092	0.2793
25	0.5436	0.5866	0.5810	0.8157	0.4973	0.3135
50	0.7135	0.7263	0.7238	0.9165	0.6301	0.3824
100	0.8324	0.8367	0.8299	0.9788	0.7551	0.4882

3.3.2 Multivariate

To evaluate clustered outcome data, values for covariates $X_{ij_1}, ..., X_{ij_{25}}$ were generated, with $X_{ij_k} = X_{i_k}$ for $k = 1, ..., 10$. For each cluster, $(\log(X_{i_1}), ..., \log(X_{i_{10}})) \sim MVN(\mathbf{0}, \mathbf{\Sigma_2})$, where Σ_2 was defined such that $corr(\log(X_{i_k}),\log(X_{i_{k'}})) = 0.5$ for $k = 1, ..., 5, k' = 1, ..., 5$ and $k = 6, ..., 10, k' = 6, ..., 10$, $corr(\log(X_{i_k}),\log(X_{i_{k'}})) = 0.2$ for $k = 1, ..., 5, k' = 6, ..., 10$. Each covariate X_{ij_k} for $k = 11, ..., 20$ was simulated from the multivariate lognormal distribution with $corr(log(X_{ij_k}), log(X_{ij'_k}))$=0.2 independently across k. Finally, for $k = 21, ..., 25$, $log(X_{ij_k}) \sim N(0, 25)$ with independence between and within clusters. Binary treatment A_i was generated with $P(A = 1) = 0.5$, with the total number of clusters assigned to each treatment level fixed accordingly. To induce unexplained correlation within clusters, random cluster effects b_i were simulated, with $log(b_i) \sim N(0, \rho\sigma^2)$, where ρ was varied to induce high or low intracluster correlation. Outcomes Y_{ij} were generated from the model $Y_{ij} = \eta_0 + \eta_1 A_i + \eta_2 X_{i_1} + \eta_3 X_{ij_{11}} + \eta_4 X_{i_3} + \eta_5 X_{ij_{12}} \eta_6 X_{ij_{15}} + b_i + \varepsilon_{ij}$, with $\log(\varepsilon_{ij}) \sim N(0, \sigma^2 = 1.9)$. We set the coefficient vector $\eta = (1, 0, 1.25, 1.25, 0.2, 0.2, 0.2)$ under the null hypothesis of no treatment effect, and $\eta = (1, 2.2, 1.25, 1.25, 0.2, 0.2, 0.2)$ under the alternative. Monte Carlo datasets consisted of $n = 10, 15, 25$ clusters of size $m_i = 20, 30$ or $n = 25, 50, 100$ clusters of size $m_i = 4, 6, 8$ per treatment arm. Values of ρ considered were $\rho = 7/19$ under the null, and $\rho = 7/19, 1$ under the alternative, corresponding to $corr(Y_{ij}, Y_{ij'}|\mathbf{X}_i, A_i)$=5% and 50%, respectively. At $\rho = 7/19$, the correlation between Y_{ij} and baseline covariates was 0.28, whereas $\rho = 1$ reduced $corr(Y_{ij}, \mathbf{X}_{ij}|A_i)$ to 0.17.

We first adaptively determined predictive models for the mean outcome conditional on baseline covariates without consideration of correlation among outcomes within a cluster. We then compared these results to the Monte Carlo power of adjusted tests when model selection accounted for correlation in responses (Section 3.2.3). Selection of baseline covariates for adjustment included forward selection by AIC, two modifications of BIC for multivariate data, and adaptive LASSO. All regression models were ultimately fit by OLS. For BIC, two regression models were selected, the first considering the number of clusters in the penalty for model complexity(BICn), and the second

calculating BIC based on the total number of individual-level observations(BICm). In deriving BIC for mixed models, Pauler (1998) showed that for a random intercept model the true penalty is of the form $\Omega_h = \sum_{k=1}^{p} log(N_k^*)$, where h indexes candidate models, k indexes the p covariates in the h^{th} model, $N_k^* = n$ for between-cluster effects, and $N_k^* = M$ for within-cluster effects. BICm and BICn therefore correspond to models containing entirely cluster-level covariates or individual-level covariates, respectively. Evaluating the true BIC for models including both types of covariates requires calculating Ω_h for each candidate model in the stepwise procedure by observing its cluster-level and individual-level covariates. To ease computation, BICm and BICn were used. The adaptive LASSO tuning parameter was selected based on five-fold cross validation. The two fixed regression models included the data generating model and an incorrect model, $E[Y_{ij}|\mathbf{X}_{ij}, A_i] = \eta_0 + \eta_1 X_{i_1} + \eta_2 X_{i_2} + \eta_3 X_{i_{10}} + \eta_4 X_{ij_{13}} + \eta_5 X_{ij_{21}}$, including two predictive covariates and 3 noisy covariates. For Methods Ib and IIb, treatment was forced into the regression model; model selection and prespecified models for the randomization tests omitted treatment. The null distribution of the observed test statistic under the exact test was determined by permuting the treatment assignment across clusters $b = 1000$ times. Unadjusted tests were also performed for each method and compared to covariate-adjusted tests. The impact of incorporating the covariance structure on randomization tests was evaluated by conducting each test under both independence and exchangeable correlation structures for each adjustment model. Specification of a covariance structure for standard GEE and augmented GEE methods have been evaluated elsewhere {Wang and Carey (2003),Stephens et al. (2012a)}.

Type I error for each method is presented in Tables 3.5-3.7. In small samples ($n_a \leq 25$) GEE methods fail to control type I error for all covariate-adjusted analyses. Inflation of type I error reflects bias in variance estimation of the sandwich estimator in small samples as well as additional variance induced by model selection. Under model selection, type I error was as large as $\alpha = 0.24$ for Method Ib and $\alpha = 0.31$ for Methods IIb. When the number of clusters was large ($n_a \geq 50$), nominal type I error levels of $\alpha = 0.05$ were achieved when covariates were not selected adaptively. Type I error

Table 3.5: **Type I Error of Multivariate AMM and Augmented tests.** Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. LASSO). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Adjusted Mean Model

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0692	0.2382	0.2100	0.1544	0.1566	0.0970	0.0958
	15	0.0596	0.1504	0.1306	0.1052	0.1040	0.0688	0.0664
	25	0.0548	0.1012	0.0946	0.0846	0.0904	0.0650	0.0676
Small m_i	25	0.0589	0.1014	0.0831	0.0779	0.0747	0.0627	0.0639
	50	0.0466	0.0642	0.0562	0.0526	0.0550	0.0470	0.0522
	100	0.0483	0.0659	0.0601	0.0607	0.0601	0.0586	0.0556

Augmented

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0692	0.3076	0.2636	0.1824	0.1982	0.1204	0.1196
	15	0.0596	0.1984	0.1650	0.1236	0.1394	0.0836	0.0838
	25	0.0548	0.1244	0.1114	0.0964	0.1128	0.0752	0.0738
Small m_i	25	0.0589	0.1234	0.0923	0.0817	0.0827	0.0710	0.0734
	50	0.0466	0.0734	0.0620	0.0578	0.0602	0.0538	0.0560
	100	0.0483	0.0665	0.0586	0.0580	0.0601	0.0601	0.0559

was still inflated under model selection for the large n considered ($\alpha = 0.05 - 0.068$ for $n_a \leq 25$), but inflation was slight compared to that observed for small n (α=0.07-0.31). For testing treatment effects, model selection by AIC resulted in the largest type I error, followed by the BIC methods; the adaptive LASSO had the least type I error inflation. For the randomization tests, the approximate exact test was generally conservative across all outcomes. The Bickel adjustment for defining the rejection region increased type I error levels of the approximate exact test closer to the nominal level. The exact test had nominal type I error across selected and prespecified covariate-adjusted models.

Table 3.6: **Type I Error of (Multivariate) Approximate Exact Tests.** Results based on Bickel's adjusted cdf are indicated by (Sm). Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. LASSO). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Approximate Exact (Ind)

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0406	0.0426	0.0384	0.0418	0.0380	0.0400	0.0438
	15	0.0460	0.0378	0.0404	0.0406	0.0392	0.0382	0.0378
	25	0.0430	0.0524	0.0514	0.0500	0.0496	0.0444	0.0484
Small m_i	25	0.0443	0.0469	0.0443	0.0451	0.0471	0.0439	0.0413
	50	0.0432	0.0408	0.0392	0.0404	0.0396	0.0386	0.0434
	100	0.0428	0.0501	0.0516	0.0531	0.0528	0.0531	0.0492

Approximate Exact (Ind-Sm)

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0412	0.0436	0.0400	0.0434	0.0392	0.0428	0.0454
	15	0.0478	0.0400	0.0416	0.0432	0.0416	0.0392	0.0390
	25	0.0444	0.0532	0.0522	0.0512	0.0516	0.0468	0.0496
Small m_i	25	0.0453	0.0479	0.0473	0.0475	0.0488	0.0455	0.0429
	50	0.0444	0.0422	0.0406	0.0414	0.0414	0.0400	0.0458
	100	0.0431	0.0519	0.0537	0.0543	0.0549	0.0549	0.0507

Approximate Exact (Exch)

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0392	0.0394	0.0384	0.0430	0.0384	0.0402	0.0434
	15	0.0432	0.0396	0.0418	0.0412	0.0402	0.0384	0.0384
	25	0.0430	0.0522	0.0518	0.0510	0.0510	0.0478	0.0480
Small m_i	25	0.0439	0.0463	0.0455	0.0453	0.0477	0.0447	0.0447
	50	0.0406	0.0412	0.0392	0.0404	0.0394	0.0390	0.0458
	100	0.0434	0.0486	0.0525	0.0525	0.0528	0.0534	0.0495

Approximate Exact (Exch-Sm)

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0394	0.0418	0.0404	0.0448	0.0402	0.0430	0.0456
	15	0.0446	0.0406	0.0430	0.0428	0.0414	0.0402	0.0390
	25	0.0440	0.0538	0.0530	0.0526	0.0528	0.0486	0.0490
Small m_i	25	0.0451	0.0481	0.0475	0.0475	0.0496	0.0467	0.0461
	50	0.0410	0.0430	0.0408	0.0418	0.0422	0.0410	0.0470
	100	0.0443	0.0504	0.0528	0.0525	0.0534	0.0537	0.0510

Table 3.7: **Type I Error of Multivariate Exact Tests.** Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. LASSO). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Exact (Ind)

	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0494	0.0496	0.0454	0.0480	0.0428	0.0490	0.0478
	15	0.0526	0.0450	0.0450	0.0466	0.0434	0.0464	0.0434
	25	0.0474	0.0568	0.0556	0.0528	0.0574	0.0510	0.0510
Small m_i	25	0.0486	0.0512	0.0492	0.0500	0.0524	0.0488	0.0498
	50	0.0466	0.0446	0.0396	0.0408	0.0420	0.0452	0.0474
	100	0.0416	0.0553	0.0543	0.0556	0.0586	0.0562	0.0522

Exact (Exch)

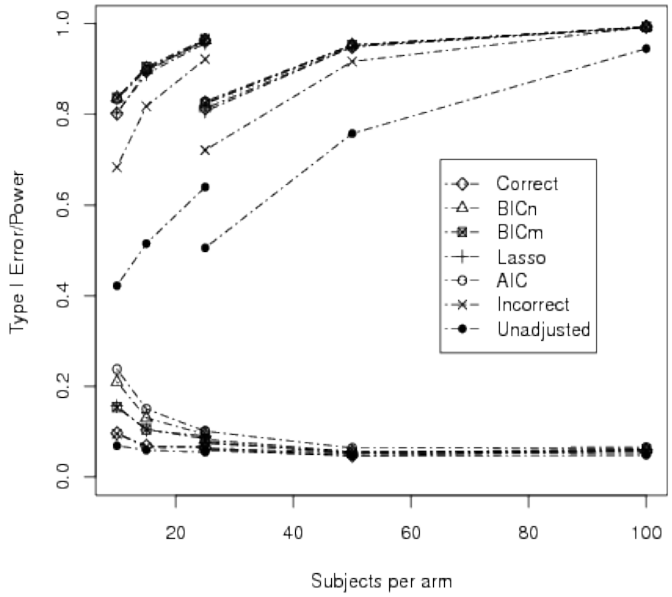
	n_a	Unadjusted	AIC	BICn	BICm	A. LASSO	Correct	Incorrect
Large m_i	10	0.0482	0.0460	0.0454	0.0486	0.0444	0.0494	0.0512
	15	0.0500	0.0470	0.0474	0.0456	0.0456	0.0464	0.0436
	25	0.0484	0.0558	0.0564	0.0560	0.0570	0.0530	0.0502
Small m_i	25	0.0481	0.0520	0.0494	0.0492	0.0522	0.0518	0.0510
	50	0.0444	0.0436	0.0408	0.0416	0.0432	0.0446	0.0476
	100	0.0464	0.0534	0.0556	0.0556	0.0580	0.0565	0.0522

Figure 3.4 and Tables 3.8-3.13 compare power across covariate-adjusted tests for dependent outcomes. In most cases, covariate adjustment improved power compared to the corresponding unadjusted approaches, regardless of the method of model selection used. Precision matrix scaling seemed to reduce overfitting in model selection; adaptive methods tended to select fewer covariates when outcomes and covariates were scaled prior to adjustment in the setting where outcomes were highly correlated (Table 3.14). Post-selection randomization tests also had larger power when outcomes and covariates were scaled before selection versus not scaled.

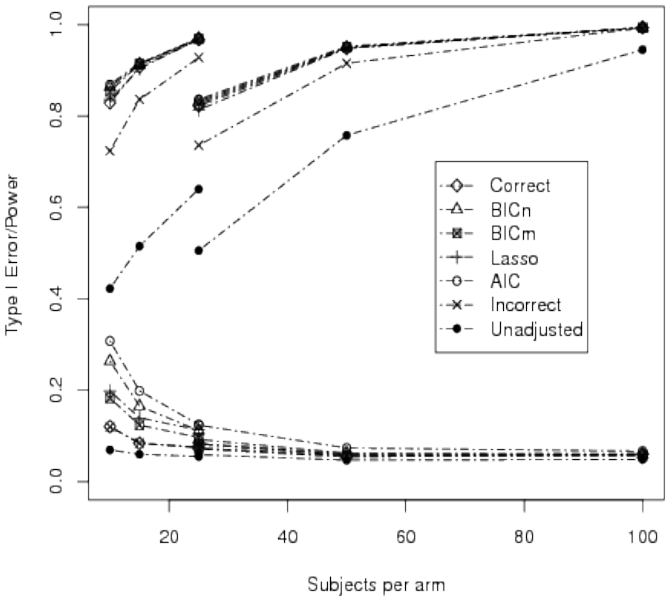
Method IVb at $n_a = 10$ AIC and BICn selection strategies had lower power than did strategies that did not adjust for baseline covariates when the exchangeable working covariance was used and precision matrix scaling was not done prior to model selection (Unadjusted 0.2170, AIC 0.1894, BICn 0.2014). Upon scaling outcomes and covariates prior to model selection, post-selection by AIC or BICn tests were more powerful than unadjusted tests (AIC 0.229, BICn 0.250). Of the adaptive methods considered, forward selection by BICm resulted in the largest power for both levels of intracluster correlation. Exchangeable working covariance specification improved power over working independence only for randomization tests of the unadjusted outcomes $\mathbf{y}_i$.

Figure 3.4: **Type I Error and Power of Multivariate AMM and Augmented Tests.** Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (Lasso). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Figure 3.4(Continued)



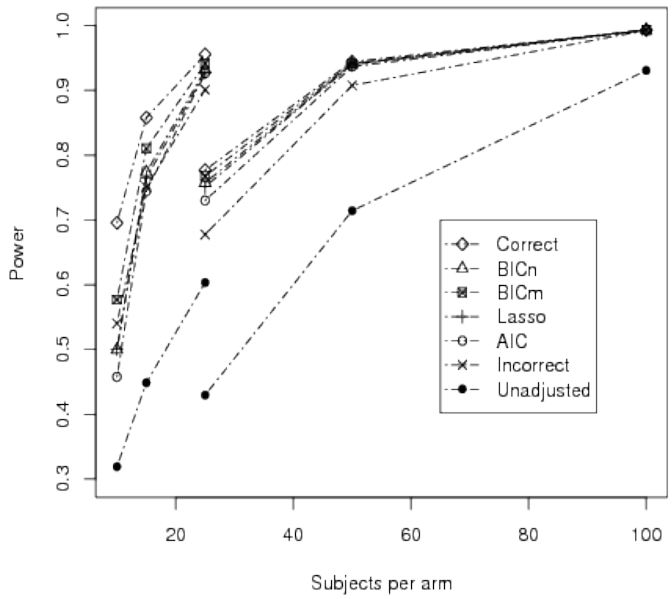
(a) AMM



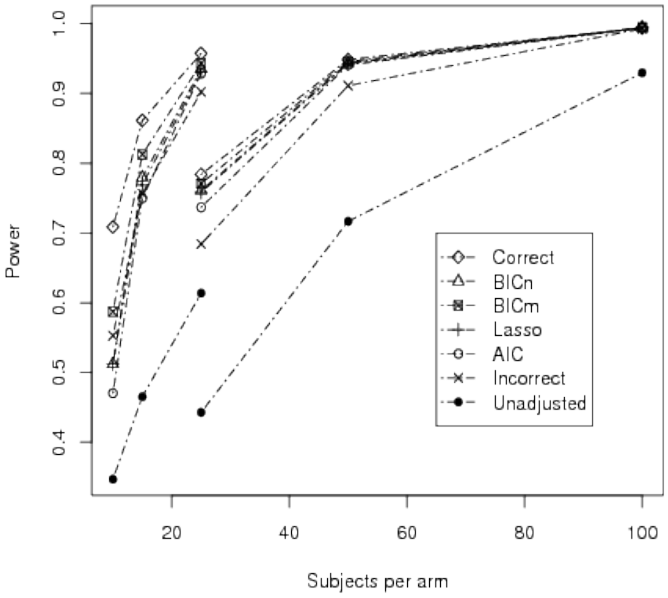
(b) Augmented

Figure 3.5: **Power of Multivariate Approx. Exact and Exact Tests: low correlation.** Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (Lasso). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Figure 3.5(Continued)



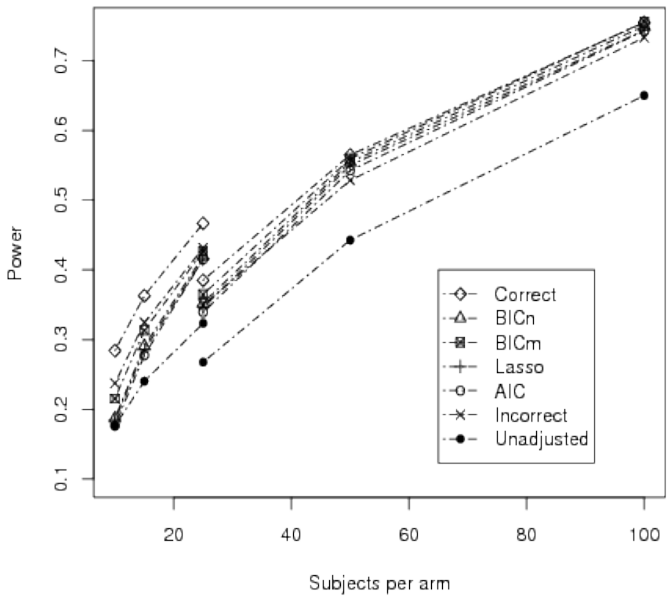
(a) Approximate



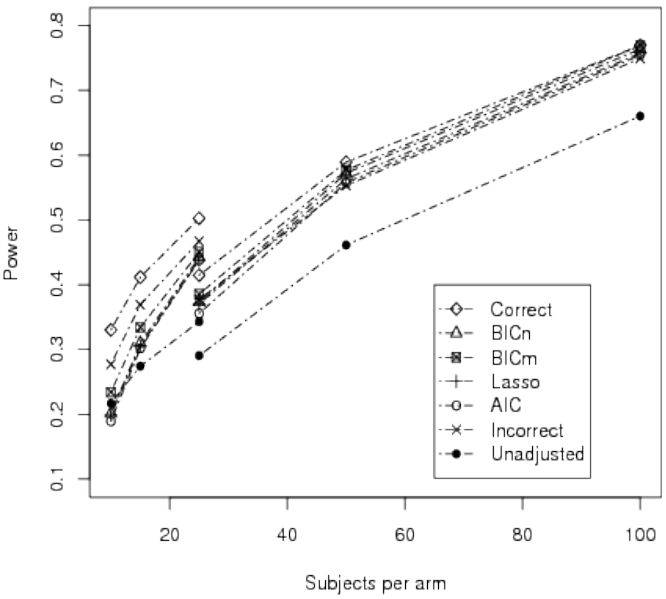
(b) Exact

Figure 3.6: **Power of Multivariate Approx. Exact and Exact Tests: high correlation.** Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (Lasso). Prespecified models: Correct, Incorrect. 'Unadjusted' denotes the test statistic that does not incorporate baseline covariates.

Figure 3.6(Continued)



(a) Approximate



(b) Exact

Table 3.8: **Power of Multivariate AMM and Augmented Tests: low correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. (*) indicates model selection on precision matrix-transformed covariates and outcomes. Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Adjusted Mean Model

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.422	0.834	0.832	0.837	0.832	0.837	0.829	0.803	0.797	0.802	0.684
15	0.515	0.899	0.901	0.901	0.903	0.905	0.905	0.889	0.884	0.895	0.818
25	0.640	0.960	0.962	0.963	0.965	0.966	0.967	0.957	0.954	0.964	0.922
25	0.505	0.829	0.830	0.825	0.826	0.823	0.822	0.806	0.806	0.813	0.721
50	0.758	0.953	0.950	0.953	0.952	0.953	0.952	0.949	0.948	0.949	0.917
100	0.945	0.993	0.994	0.993	0.993	0.993	0.993	0.992	0.993	0.994	0.993

Augmented

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.422	0.869	0.863	0.863	0.858	0.854	0.847	0.836	0.833	0.829	0.724
15	0.515	0.915	0.914	0.914	0.915	0.914	0.912	0.904	0.905	0.907	0.836
25	0.640	0.970	0.969	0.970	0.970	0.969	0.968	0.968	0.965	0.967	0.928
25	0.505	0.836	0.837	0.832	0.832	0.827	0.824	0.813	0.812	0.822	0.736
50	0.758	0.953	0.952	0.950	0.949	0.950	0.950	0.948	0.947	0.948	0.915
100	0.945	0.994	0.994	0.993	0.994	0.993	0.994	0.993	0.994	0.993	0.993

Table 3.9: **Power of Multivariate Approximate Exact Tests: low correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. (*) indicates model selection on precision matrix-transformed covariates and outcomes. Results based on Bickel's adjusted CDF are indicated by (Sm). Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Approximate Exact (Ind)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.221	0.453	0.482	0.496	0.530	0.566	0.604	0.495	0.494	0.686	0.529
15	0.325	0.740	0.777	0.769	0.806	0.802	0.829	0.759	0.762	0.853	0.738
25	0.465	0.923	0.935	0.930	0.943	0.939	0.948	0.925	0.927	0.952	0.897
25	0.322	0.725	0.735	0.754	0.760	0.763	0.766	0.748	0.748	0.769	0.671
50	0.564	0.933	0.935	0.938	0.939	0.939	0.939	0.935	0.937	0.941	0.905
100	0.827	0.992	0.993	0.992	0.993	0.993	0.993	0.993	0.993	0.993	0.992

Approximate Exact (Ind-Sm)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.226	0.460	0.491	0.503	0.539	0.574	0.612	0.501	0.500	0.692	0.536
15	0.328	0.744	0.780	0.773	0.810	0.807	0.833	0.764	0.768	0.856	0.743
25	0.467	0.925	0.937	0.931	0.944	0.940	0.949	0.926	0.928	0.953	0.901
25	0.326	0.730	0.741	0.759	0.767	0.769	0.772	0.753	0.752	0.776	0.675
50	0.568	0.936	0.938	0.941	0.942	0.943	0.942	0.938	0.940	0.943	0.907
100	0.831	0.993	0.994	0.993	0.994	0.993	0.994	0.994	0.994	0.994	0.992

Approximate Exact (Exch)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.315	0.450	0.484	0.493	0.531	0.568	0.606	0.493	0.495	0.690	0.536
15	0.445	0.739	0.777	0.769	0.809	0.806	0.832	0.760	0.763	0.855	0.748
25	0.602	0.925	0.938	0.930	0.944	0.940	0.950	0.927	0.927	0.955	0.898
25	0.425	0.726	0.733	0.753	0.760	0.762	0.766	0.746	0.747	0.771	0.674
50	0.712	0.935	0.936	0.937	0.939	0.939	0.940	0.937	0.937	0.942	0.906
100	0.930	0.992	0.994	0.993	0.994	0.993	0.994	0.993	0.994	0.993	0.992

Approximate Exact (Exch-Sm)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.319	0.458	0.489	0.500	0.539	0.577	0.613	0.501	0.502	0.696	0.540
15	0.449	0.744	0.781	0.774	0.812	0.810	0.837	0.764	0.768	0.858	0.751
25	0.604	0.927	0.940	0.932	0.946	0.941	0.951	0.928	0.929	0.956	0.901
25	0.430	0.730	0.739	0.758	0.766	0.768	0.772	0.751	0.752	0.777	0.678
50	0.714	0.937	0.938	0.940	0.941	0.942	0.942	0.941	0.940	0.945	0.908
100	0.931	0.993	0.994	0.993	0.994	0.994	0.994	0.994	0.994	0.993	0.993

Table 3.10: **Power of Multivariate Exact Tests: low correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Exact (Ind)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.246	0.472	0.502	0.512	0.548	0.587	0.622	0.510	0.513	0.705	0.550
15	0.338	0.751	0.785	0.776	0.815	0.811	0.836	0.767	0.770	0.862	0.751
25	0.473	0.927	0.938	0.934	0.945	0.940	0.950	0.929	0.929	0.956	0.902
25	0.335	0.735	0.744	0.763	0.771	0.773	0.776	0.759	0.759	0.785	0.681
50	0.570	0.940	0.940	0.943	0.944	0.944	0.944	0.942	0.943	0.948	0.909
100	0.830	0.994	0.995	0.994	0.995	0.995	0.995	0.995	0.995	0.994	0.992

Exact (Exch)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.347	0.470	0.504	0.512	0.550	0.587	0.622	0.512	0.515	0.709	0.553
15	0.465	0.750	0.785	0.780	0.817	0.813	0.837	0.769	0.772	0.861	0.756
25	0.614	0.929	0.940	0.935	0.948	0.943	0.953	0.931	0.931	0.957	0.902
25	0.443	0.737	0.744	0.761	0.771	0.771	0.774	0.758	0.758	0.784	0.684
50	0.717	0.941	0.941	0.944	0.945	0.946	0.946	0.943	0.943	0.949	0.911
100	0.930	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995	0.994	0.993

Table 3.11: **Power of Multivariate AMM and Augmented tests: high correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. (*) indicates model selection on precision matrix-transformed covariates and outcomes. Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Adjusted Mean Model

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.252	0.524	0.482	0.527	0.477	0.519	0.467	0.503	0.487	0.409	0.356
15	0.297	0.511	0.486	0.515	0.485	0.514	0.479	0.498	0.491	0.449	0.412
25	0.350	0.544	0.532	0.547	0.531	0.549	0.526	0.537	0.527	0.504	0.477
25	0.308	0.487	0.470	0.477	0.462	0.468	0.455	0.459	0.448	0.431	0.395
50	0.466	0.611	0.606	0.605	0.603	0.604	0.603	0.600	0.599	0.590	0.558
100	0.663	0.768	0.769	0.771	0.766	0.770	0.766	0.769	0.761	0.765	0.742

Augmented

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.252	0.630	0.547	0.527	0.459	0.493	0.434	0.607	0.566	0.449	0.401
15	0.297	0.591	0.523	0.507	0.481	0.491	0.465	0.575	0.544	0.479	0.442
25	0.350	0.583	0.551	0.539	0.533	0.532	0.523	0.578	0.557	0.524	0.488
25	0.308	0.515	0.486	0.470	0.463	0.462	0.455	0.487	0.467	0.453	0.414
50	0.466	0.623	0.608	0.602	0.603	0.602	0.601	0.613	0.605	0.598	0.563
100	0.663	0.771	0.767	0.766	0.769	0.763	0.767	0.770	0.766	0.764	0.745

Table 3.12: **Power of Multivariate Approximate Exact Tests: high correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. (*) indicates model selection on precision matrix-transformed covariates and outcomes. Results based on Bickel's adjusted CDF are indicated by (Sm). Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Approximate Exact (Ind)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.140	0.170	0.200	0.181	0.217	0.211	0.248	0.181	0.191	0.278	0.232
15	0.197	0.270	0.316	0.280	0.328	0.306	0.340	0.279	0.299	0.355	0.322
25	0.268	0.411	0.438	0.414	0.453	0.421	0.458	0.412	0.422	0.463	0.430
25	0.213	0.328	0.351	0.342	0.365	0.352	0.367	0.340	0.344	0.375	0.342
50	0.355	0.532	0.545	0.541	0.554	0.547	0.557	0.536	0.542	0.554	0.522
100	0.557	0.733	0.744	0.740	0.749	0.743	0.749	0.734	0.736	0.744	0.717

Approximate Exact (Ind-Sm)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.142	0.173	0.205	0.185	0.223	0.216	0.252	0.185	0.194	0.284	0.235
15	0.197	0.274	0.320	0.284	0.333	0.310	0.344	0.281	0.303	0.359	0.324
25	0.270	0.413	0.441	0.417	0.454	0.423	0.460	0.415	0.425	0.466	0.432
25	0.215	0.332	0.354	0.345	0.369	0.356	0.371	0.344	0.349	0.379	0.344
50	0.357	0.535	0.549	0.546	0.558	0.550	0.561	0.540	0.548	0.558	0.525
100	0.559	0.734	0.746	0.740	0.751	0.744	0.751	0.736	0.738	0.746	0.719

Approximate Exact (Exch)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.174	0.172	0.198	0.183	0.215	0.212	0.247	0.181	0.191	0.281	0.234
15	0.239	0.274	0.320	0.287	0.333	0.311	0.345	0.283	0.300	0.359	0.322
25	0.321	0.413	0.443	0.416	0.453	0.423	0.458	0.413	0.427	0.466	0.430
25	0.266	0.334	0.356	0.348	0.371	0.360	0.374	0.341	0.350	0.380	0.346
50	0.442	0.538	0.553	0.550	0.562	0.556	0.563	0.546	0.548	0.561	0.526
100	0.649	0.740	0.748	0.748	0.751	0.754	0.752	0.742	0.742	0.753	0.732

Approximate Exact (Exch-Sm)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.176	0.177	0.201	0.187	0.219	0.215	0.252	0.184	0.193	0.284	0.237
15	0.240	0.278	0.323	0.291	0.337	0.313	0.348	0.286	0.304	0.363	0.325
25	0.323	0.415	0.445	0.419	0.456	0.426	0.459	0.415	0.430	0.467	0.431
25	0.268	0.339	0.360	0.353	0.374	0.365	0.377	0.346	0.354	0.385	0.349
50	0.443	0.542	0.558	0.554	0.565	0.559	0.566	0.550	0.552	0.565	0.528
100	0.650	0.743	0.751	0.750	0.752	0.755	0.754	0.744	0.746	0.755	0.733

Table 3.13: **Power of Multivariate Exact Tests: high correlation.** Rows 1-3 contain results for cluster size $m_i = (20, 30)$. Rows 4-6 show results for $m_i = (4, 6, 8)$. Adaptive regression model selection: AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. L.). Prespecified models: Correct (Corr.), Incorrect (Inco.). 'Unadj.' denotes the test statistic that does not incorporate baseline covariates.

Exact (Ind)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.175	0.188	0.231	0.200	0.250	0.234	0.283	0.199	0.215	0.330	0.275
15	0.222	0.294	0.347	0.304	0.367	0.331	0.380	0.302	0.330	0.410	0.366
25	0.295	0.436	0.472	0.442	0.483	0.450	0.487	0.437	0.451	0.505	0.463
25	0.231	0.351	0.379	0.365	0.392	0.378	0.398	0.367	0.374	0.409	0.370
50	0.369	0.552	0.571	0.566	0.579	0.569	0.578	0.556	0.565	0.583	0.546
100	0.571	0.748	0.758	0.754	0.762	0.759	0.761	0.752	0.752	0.761	0.732

Exact (Exch)

n_a	Unadj.	AIC	AIC*	BICn	BICn*	BICm	BICm*	A. L.	A. L.*	Corr.	Inco.
10	0.217	0.189	0.229	0.201	0.250	0.234	0.284	0.199	0.215	0.330	0.277
15	0.274	0.302	0.355	0.310	0.372	0.334	0.385	0.305	0.329	0.412	0.369
25	0.343	0.438	0.474	0.442	0.483	0.450	0.490	0.439	0.457	0.503	0.467
25	0.291	0.356	0.384	0.373	0.398	0.386	0.399	0.370	0.380	0.415	0.377
50	0.461	0.558	0.574	0.572	0.585	0.577	0.587	0.564	0.572	0.589	0.553
100	0.661	0.755	0.762	0.764	0.768	0.769	0.768	0.758	0.758	0.770	0.749

Table 3.14: **Average Number of Baseline Covariates** selected by AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. LASSO) by sample size when outcomes were multivariate. Rows 1-3 contain results for cluster size $m_i = (20, 30)$; rows 4-6 for $m_i = (4, 6, 8)$. Results are shown for estimating $E[\mathbf{Y}_i|\mathbf{X}_i]$ considering untransformed (U) and transformed (T) covariates and outcomes.

Low Correlation

	AIC		BICn		BICm		A. LASSO	
n_a	U	T	U	T	U	T	U	T
10	8.61	9.55	6.65	7.67	3.95	5.12	7.66	8.57
15	9.02	9.33	6.48	7.05	4.10	4.98	8.22	8.79
25	9.45	9.62	6.37	7.01	4.29	5.31	8.77	9.51
25	6.84	7.65	4.11	5.08	3.13	4.15	4.51	5.28
50	7.27	7.96	3.98	4.98	3.22	4.28	4.86	5.52
100	7.82	8.49	4.23	5.28	3.55	4.67	5.93	6.50

High Correlation

	Aic		BICn		BICm		Adap Lasso	
n_a	U	T	U	T	U	T	U	T
10	10.93	9.70	8.95	7.79	5.84	5.24	11.52	9.87
15	11.30	9.44	8.76	7.28	5.99	5.26	12.34	9.65
25	11.69	9.69	8.53	7.30	6.06	5.70	13.01	9.74
25	8.06	7.81	4.99	5.35	3.70	4.41	6.81	5.70
50	8.51	8.61	4.72	5.73	3.72	4.92	7.31	5.94
100	8.86	9.67	4.66	6.46	3.80	5.75	7.94	6.43

3.4 Application

Covariate-adjusted tests were applied to data from the *Young Citizens* study. *Young Citizens* was a cluster-randomized intervention trial designed to evaluate the effectiveness of a behavioral intervention in training adolescents to be peer educators about HIV. Thirty communities were randomized to intervention or control, resulting in 15 communities per arm. Residents in participating communities were surveyed regarding the degree to which they believed adolescents could effectively communicate to their families and peers about HIV transmission dynamics. The outcome Y_{ij} was a child empowerment score from responses given by individuals within each randomized community. Additional covariates characterizing the communities and households of survey respondents were measured.

Predictive models for baseline covariates were first determined by AIC, BICn, BICm, and adaptive LASSO. Covariates selected by AIC include employment status (employment), age of the head of household (age), whether or not the household had a flushing toilet (flushing toilet), number of relatives in the neighborhood (relatives), religion, community population density (density), transportation ownership (transportation), home ownership (home), and interactions of treatment with relatives and density. BICn selected the same covariates as AIC except for transportation and home, which it did not enter into the model. BIC penalized by the number of total observations (BICm) chose employment, age, and flushing toilet. Finally, adaptive LASSO picked flushing toilet, religion, employment, age, and interactions with treatment and density, relatives, and number of kids in the house. For randomization tests, the AIC-based model contained employment, flushing toilet, age, religion, relatives, home, and wealth deviance for each family from the mean community wealth. BICn selected employment, flushing toilet, age, religion and relatives. Selection by BICm and adaptive LASSO chose employment, flushing toilet, and age.

Table 3.15: **Analysis of the** *Young Citizens* **study.** Covariate-adjusted method (Method), regression (OR) {AIC, BIC by n (BICn), BIC by M,(BICm), Adaptive LASSO (A. LASSO)}, test statistic (T) and p-value (p), with each test statistic evaluated under independence (Ind) and exchangeable (Exch) working covariance. P-values for Approx. Exact tests are calculated under Bickel's cdf for randomization test statistics. 'Unadjusted' denotes the unadjusted test.

		Ind		Exch	
Method	OR	Test Stat	p	Test Stat	p
Adjusted	AIC	58.7003	< 0.0001	53.5700	< 0.0001
	BICm	59.9557	< 0.0001	54.5695	< 0.0001
	BICm	4.5046	< 0.0001	4.6231	< 0.0001
	A. LASSO	112.0423	< 0.0001	103.4147	< 0.0001
	Unadjusted	4.1415	< 0.0001	4.3186	< 0.0001
Augmented	AIC	5.1136	< 0.0001	5.2477	< 0.0001
	BICM	5.1845	< 0.0001	5.2321	< 0.0001
	BICN	4.6400	< 0.0001	4.6565	< 0.0001
	Adaptive LASSO	5.3805	< 0.0001	5.3756	< 0.0001
Approx. Exact	AIC	3.1326	0.0017	3.3316	0.0009
	BICm	3.1431	0.0017	3.3836	0.0007
	BICn	3.1223	0.0018	3.3280	0.0009
	A. LASSO	3.1223	0.0018	3.3280	0.0009
	Unadjusted	1.6172	0.1058	2.2682	0.0233
Approx. Exact (Sm)	AIC	3.1326	0.0017	3.3316	0.0008
	BICm	3.1431	0.0017	3.3836	0.0007
	BICn	3.1223	0.0018	3.3280	0.0009
	A. LASSO	3.1223	0.0018	3.3280	0.0009
	Unadjusted	1.6170	0.1060	2.2682	0.0233
Exact	AIC	89.8329	0.0003	37.0575	0.0003
	BICm	91.9124	0.0007	36.5084	0.0003
	BICn	88.8094	0.0007	36.5876	0.0007
	A. LASSO	88.8094	0.0007	26.5876	0.0007
	Unadjusted	434.8410	0.1043	71.4085	0.1200

Table 3.15 presents results from the *Young Citizens* analysis. Adjusted and augmented GEE methods were associated with highly significant treatment effects ($p < 0.0001$) across covariate-adjusted and unadjusted tests. For the approximate exact tests, all covariate-adjusted methods yielded a significant intervention effect. When unadjusted, however, only the test using exchangeable covariance resulted in significantly different child empowerment between intervention groups ($p = 0.0233$ for exchangeable working covariance, $p = 0.10$ under independence). Applying Bickel's small-sample adjustment to obtain tail probabilities resulted in p-values that were slightly larger than those based on the standard normal distribution. Among permutation tests, significant intervention effects were detected under covariate-adjustment, but not in the absence of such adjustment for either working covariance structure. The data provide sufficient evidence that children who participated in the intervention were significantly more equipped to educate their peers about HIV. The results underscore the importance of using appropriate methodology and utilizing baseline covariate information. Unadjusted tests based on GEE methods were highly significant, but with a fairly small number of clusters, the validity of such methods is not guaranteed. Randomization tests, with guaranteed validity in small samples, showed similar results with covariate adjustment, but conclusions of unadjusted tests were inconsistent.

3.5 Discussion

We have investigated the dangers and merits of several procedures that allow for flexible covariate adjustment when applied to small samples. Simulation studies showed, as expected, that AMM and augmented methods break down in small samples when the number of baseline covariates is large relative to the sample size. Alternatively, randomization methods, which exploit the fact that outcomes and baseline covariates are regarded as fixed, provide valid tests for treatment effects when flexibly incorporating baseline covariates. Model selection may be used to identify the set of baseline covariates that explain the greatest amount of variability in the outcome while preserving the type I error of the primary test. The central conclusion is that randomization tests therefore do not require adjustment models to be prespecified to preserve the nominal type I error. Furthermore, adjustment generally increases the power of testing for treatment effects over unadjusted methods, with the caveat that in extremely small samples of independent outcomes, such as $n_a = 10, 15$, model selection approaches must be sufficiently conservative. Model selection by BIC and adaptive LASSO, which have stronger penalties and therefore favor more parsimonious models than AIC, resulted in improved power at the smallest sample sizes considered. Further research is needed to formally characterize the power of covariate-adjusted tests under misspecified covariate adjustment and adaptive covariate selection.

Our presentation has focused on hypothesis testing for evaluating treatment effects. For confidence interval estimation, hypothesis tests may be inverted. When inverting randomization-based hypothesis tests, it is important to note that for each potential value of the treatment effect considered, adaptive selection needs to be repeated, since conditional mean models are estimated by pooling across treated and untreated subjects. Interval estimation may be simplified by a slight modification of the testing procedure. Under the null, the conditional mean model may be estimated using data only for untreated subjects. The model may then be applied to all subjects in conducting the test. Not pooling the data when estimating the conditional mean model removes the need for

its re-estimation with each treatment effect value considered. For small-sample univariate data, it may not be feasible to perform model selection on one treatment group may be infeasible, but for a small number of moderately sized clusters such a strategy may be more reasonable.

Appendix

Appendix A: $\hat{\beta}$ Solutions for Standard and Augmented Logistic GEE in cluster randomized designs

Let Y_{ij} denote the response (0 or 1) for the j_{th} individual in the i_{th} cluster. $\mathbf{Y}_i = (Y_{i1},\ Y_{i2}, ..., Y_{in_i})^{\mathrm{T}}$, where n_i is the number of subjects within the i_{th} cluster. A typical model for binary data is $E(Y_{ij}|A_i) = g(A_i;\beta) = g(\beta_0 + \beta_1 A_i)$, where g is the inverse logit link function. The Standard GEE for the marginal treatment effect are given by

$$\sum_{i=1}^{m} \psi_i(\mathbf{Y}, A; \beta) = \sum_{i=1}^{m} \mathbf{D}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \mathbf{g}(A_i;\beta)\} = \mathbf{0}, \tag{4.1}$$

where bold $\mathbf{g}(A_i;\beta)$ denotes the n_i-dimensional link function for the outcome vector $\mathbf{Y}_i$, $\mathbf{D}_i$ is the $n_i \times p$ matrix defined by $\frac{\partial \mathbf{g}(A_i;\beta)}{\partial \beta^{\mathrm{T}}}$, and $\mathbf{V}_i$ is a $n_i \times n_i$ working covariance matrix for $\mathbf{Y}_i$.

$\mathbf{D}_i$ is composed of the n_i-dimensional columns $\vec{D}_{i.0} = \frac{\partial \mathbf{g}(A_i;\beta)}{\partial \beta_0}$ and $\vec{D}_{i.1} = \frac{\partial \mathbf{g}(A_i;\beta)}{\partial \beta_1}$. Because of the cluster-randomized design, $\vec{D}_{i.0}$ and $\vec{D}_{i.1}$ are vectors of the form $\vec{D}_{i.p} = (D_{i_p}, D_{i_p}, \ldots, D_{i_p})^{\mathrm{T}}$ for $p = 0, 1$ (intercept and treatment effect), with

$$D_{i_0} = \frac{\partial g(A_i;\beta)}{\partial \beta_0} = \frac{\exp(\beta_0 + \beta_1 A_i)}{(1 + \exp(\beta_0 + \beta_1 A_i))} \left(1 - \frac{\exp(\beta_0 + \beta_1 A_i)}{(1 + \exp(\beta_0 + \beta_1 A_i))}\right) = \pi(A_i)\{1 - \pi(A_i)\} \tag{4.2}$$

$$D_{i_1} = \frac{\partial g(A_i;\beta)}{\partial \beta_1} = \frac{\exp(\beta_0 + \beta_1 A_i) A_i}{(1 + \exp(\beta_0 + \beta_1 A_i))} \left(1 - \frac{\exp(\beta_0 + \beta_1 A_i) A_i}{(1 + \exp(\beta_0 + \beta_1 A_i))}\right) = \pi(A_i)\{1 - \pi(A_i)\} A_i,$$

where $\pi(A_i) = E(Y_{ij}|A_i)$. We recall that $\mathbf{D}_i$ is evaluated using an initial estimator $\hat{\beta}_{init}$, usually obtained from standard logistic regression that does not account for clustering.

The inverse working covariance matrix $\mathbf{V}_i^{-1}$ can be broken down into its columns and scalar elements. Let $\mathbf{V}_i^{-1} = \left[\begin{array}{cccc} \vec{V}_{i.1}^{-1} & \vec{V}_{i.2}^{-1} & \cdots & \vec{V}_{i.n_i}^{-1} \end{array}\right]$, where $\vec{V}_{i.j}^{-1}$ is the j_{th} column of $\mathbf{V}_i^{-1}$, and $V_{i_{q,j}}^{-1}$ represents the scalar element in the q_{th} row, j_{th} column. Using this construction, after some matrix algebra, a closed form solution for β under a cluster randomized

design is given by

$$\hat{\beta}_0 = logit\left(\left[\sum_{i=1}^{n}\left\{I(A_i=0)D_{i_0}\sum_{q,j\leq n_i}V^{-1}_{i_{q,j}}\right\}\right]^{-1}\left[\sum_{i=1}^{n}\left\{(\vec{D}_{i\cdot 0}-\vec{D}_{i\cdot 1})^{\mathrm{T}}\sum_{j=1}^{n_i}(\vec{V}^{-1}_{i\cdot j}Y_{ij})\right\}\right]\right) \tag{4.3}$$

$$\hat{\beta}_1 = logit\left(\left[\sum_{i=1}^{n}\left\{I(A_i=1)D_{i_1}\sum_{q,j\leq n_i}V^{-1}_{i_{q,j}}\right\}\right]^{-1}\left[\sum_{i=1}^{n}\left\{\vec{D}^{\mathrm{T}}_{i\cdot 1}\sum_{j=1}^{n_i}(\vec{V}^{-1}_{i\cdot j}Y_{ij})\right\}\right]\right)-\hat{\beta}_0$$

This solution can be simplified using the working covariance structure. Under exchangeable correlation, $V_{i_{q,q}} = \phi$ and $V_{i_{q,j}} = \rho$ for $q \neq j$. We note working independence as a special case with off-diagonal elements $\rho = 0$. Proceeding, let ϕ^{-1} and ρ^{-1} denote the diagonal and off-diagonal elements of $\mathbf{V}_i^{-1}$, respectively. The above simplifies to

$$\begin{aligned}\hat{\beta}_0 = logit\Bigg(&\left[\sum_{i=1}^{n}D_{i_0}I(A_i=0)\left\{n_i\phi^{-1}+n_i(n_i-1)\rho^{-1}\right\}\right]^{-1}\times\\ &\left[\sum_{i=1}^{n}\left\{(D_{i_0}-D_{i_1})\{(n_i-1)\rho_i^{-1}+\phi^{-1}\}\sum_{j=1}^{n_i}Y_{ij}\right\}\right]\Bigg)\\ \hat{\beta}_1 = logit\Bigg(&\left[\sum_{i=1}^{n}I(A_i=1)D_{i_1}\{n_i\phi^{-1}+n_i(n_i-1)\rho_1^{-1}\}\right]^{-1}\times\\ &\left[\sum_{i=1}^{n}\left\{D_{i_1}\{(n_i-1)\rho^{-1}+\phi^{-1}\}\sum_{j=1}^{n_i}Y_{ij}\right\}\right]\Bigg)-\hat{\beta}_0\end{aligned} \tag{4.4}$$

In the case of the augmented GEE, we estimate $\hat{\beta}$ using the augmented estimating equations

$$\sum_{i=1}^{m}\left[\mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\left\{\mathbf{Y}_i-\mathbf{g}(A_i;\beta)\right\}-(A_i-\pi)\hat{\gamma}(\mathbf{X}_i)\right]=\mathbf{0} \tag{4.5}$$

where $\hat{\gamma}(X_i) = \left[\mathbf{D}_i(1)^{\mathrm{T}}\mathbf{V}_i(1)^{-1}\left\{f_1(\mathbf{X}_i;\hat{\eta}_1)-\mathbf{g}(1;\beta)\right\}-\mathbf{D}_i(0)^{\mathrm{T}}\mathbf{V}_i(0)^{-1}\left\{f_0(\mathbf{X}_i;\hat{\eta}_0)-\mathbf{g}(0;\beta)\right\}\right]$.

Above, we take $\mathbf{D}_i(k) = \frac{\partial \mathbf{g}(k;\beta)}{\partial \beta^{\mathrm{T}}}$, $\mathbf{V}_i(k) = \mathbf{V}_i$ evaluated under treatment k, and $f_k(\mathbf{X}_i;\hat{\eta}_k) = \hat{E}[\mathbf{Y}_i|A_i = k, \mathbf{X}_i]$ for $k = 0, 1$. Vectors $\vec{D}_{i_p}(k)$, $\vec{V}^{-1}_{i\cdot j}(k)$, and scalars $D_{i_p}(k)$, $V^{-1}_{q,j}(k)$ are defined similarly as above, evaluated under treatment k. For brevity, we write $\mathbf{F}_{k_i}$ below, where $\mathbf{F}_{k_i}$=$f_k(\mathbf{X}_i;\hat{\eta}_k)$. Solutions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$\hat{\beta}_0 = logit\Bigg(\Bigg[(1-\pi)\sum_{i=1}^{n}\Bigg\{D_{i_0}(0)\sum_{q,j\leq n_i}V_{i_{q,j}}^{-1}(0)\Bigg\}\Bigg]^{-1}\times$$
$$\sum_{i=1}^{n}\Bigg[\sum_{j=1}^{n_i}\Big\{(\vec{D}_{i\cdot_0}-\vec{D}_{i\cdot_1})^{\mathrm{T}}V_{i\cdot j}^{-1}Y_{ij}-(A_i-\pi)\Big(-\vec{D}_{i\cdot_0}(0)\vec{V}_{i\cdot j}^{-1}(0)F_{0_{ij}}\Big)\Big\}\Bigg]\Bigg) \quad (4.6)$$
$$\hat{\beta}_1 = logit\Bigg(\Bigg[\pi\sum_{i=1}^{n}\Bigg\{D_{i_1}(1)\sum_{q,j\leq n_i}V_{i_{q,j}}^{-1}(1)\Bigg\}\Bigg]^{-1}\times$$
$$\sum_{i=1}^{n}\Bigg[\sum_{j=1}^{n_i}\Big\{(\vec{D}_{i\cdot_1}^{\mathrm{T}}V_{i\cdot j}^{-1}Y_{ij})-(A_i-\pi)\Big(\vec{D}_{i\cdot_1}(1)^{\mathrm{T}}\vec{V}_{i\cdot j}^{-1}(1)F_{1_{ij}}\Big)\Big\}\Bigg]\Bigg)-\hat{\beta}_0,$$

The simplified expression in case of exchangeable structure is

$$\hat{\beta}_0 = logit\Bigg(\Bigg[(1-\pi)\sum_{i=1}^{n}D_{i_0}(0)\{n_i\phi^{-1}+n_i(n_i-1)\rho_0^{-1}\}\Bigg]^{-1}\times$$
$$\sum_{i=1}^{n}\Bigg[(D_{i_0}-D_{i_1})\{(n_i-1)\rho_i^{-1}+\phi_i^{-1}\}\sum_{j=1}^{n_i}Y_{ij}-$$
$$(A_i-\pi)\Bigg\{-D_{i_1}(0)\{(n_i-1)\rho_0^{-1}+\phi_0^{-1}\}\sum_{j=1}^{n_i}F_{0_{ij}}\Bigg\}\Bigg]\Bigg) \quad (4.7)$$
$$\hat{\beta}_1 = logit\Bigg(\Bigg[\pi\sum_{i=1}^{n}\Bigg\{D_{i_1}(1)\{n_i\phi_1^{-1}+n_i(n_i-1)\rho_1^{-1}\}\Bigg\}\Bigg]^{-1}\times$$
$$\sum_{i=1}^{n}\Bigg[D_{i_1}\{(n_i-1)\rho_i^{-1}+\phi_i^{-1}\}\sum_{j=1}^{n_i}Y_{ij}-$$
$$(A_i-\pi)\Bigg\{D_{i_1}(1)\{(n_i-1)\rho_1^{-1}+\phi_1^{-1}\}\sum_{j=1}^{n_i}F_{1_{ij}}\Bigg\}\Bigg]\Bigg)-\hat{\beta}_0,$$

where we maintian the index i in $D_{i_p}(k)$ to be consistent with the unsimplified expressions above, in which the index i on $\vec{D}_{i\cdot_p}(k)$ is retained to be mindful of varying cluster size. The quantity $D_{i_p}(k)$, however, is a fixed function of $E(Y_{ij}|A_i=k)$.

Appendix B: Variance Estimators

Let $\mathbf{Y}_i$ be the n_i-dimensional response vector, A_i the scalar treatment variable, and $\mathbf{X}_i$ a collection of baseline covariates potentially at the cluster and individual level. The

model $E(\mathbf{Y}_i|A_i) = \mathbf{g}(A_i;\beta)$ is assumed, and the estimator $\hat{\beta}$ is obtained by solving the augmented estimating equations detailed in Section 1.2. Recall that $\mathbf{V_i}$ is a working covariance matrix as typically used in GEE for estimating coefficients in restricted moment models and π=$P(A_i = 1)$. Formulas for the variance estimators discussed in Section 1.3 are presented below. The asymptotic variability of $\hat{\beta}_{aug}$ is shown to be $var(\hat{\beta}_{aug}) = \Gamma^{-1}\Delta\Gamma^{-1^{\mathrm{T}}}$, where $\Gamma = E\left[\frac{\partial\psi_{i_{opt}}(\mathbf{Y},A,\mathbf{X};\beta)}{\partial\beta^{\mathrm{T}}}\right]$, and $\Delta = E\left[\psi_{i_{opt}}(\mathbf{Y},A,\mathbf{X};\beta)^{\otimes 2}\right]$, with $U^{\otimes 2} = UU^{\mathrm{T}}$. In each of the below, $\hat{\Gamma} = m^{-1}\sum_i \mathbf{D_i}^{\mathrm{T}}\mathbf{V_i}^{-1}\mathbf{D_i}$. The four variance estimators considered are:

1. $\hat{var}_1(\hat{\beta}_{aug}) = \hat{\Gamma}^{-1}\hat{\Delta}\hat{\Gamma}^{-1^{\mathrm{T}}}$, where $\hat{\Delta} = m^{-1}\sum_{i=1}^{n}\hat{\psi}_{i_{opt}}^{\otimes 2}$, and

   $\hat{\psi}_{i_{opt}}(\mathbf{Y},A,\mathbf{X};\beta) = \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\{\mathbf{Y}_i - \mathbf{g}(A_i;\hat{\beta}_{aug})\} - (A_i - \pi)\times$
   $\left[\mathbf{D}_i(1)^{\mathrm{T}}\mathbf{V}_i(1)^{-1}\{f_1(\mathbf{X}_i;\hat{\eta}_1) - \mathbf{g}(1;\hat{\beta}_{aug})\} - \mathbf{D}_i(0)^{\mathrm{T}}\mathbf{V}_i(0)^{-1}\{f_0(\mathbf{X}_i;\hat{\eta}_0) - \mathbf{g}(0;\hat{\beta}_{aug})\}\right]$

2. $\hat{var}_2(\hat{\beta}_{aug}) = \hat{\Gamma}^{-1}\hat{\Delta}^{*}\hat{\Gamma}^{-1^{\mathrm{T}}}$, where $\hat{\Delta}^{*} = m^{-1}\sum_{i=1}^{n}(\mathbf{H}_i\hat{\psi}_i)^{\otimes 2}$, and $\mathbf{H}_i$ is a diagonal matrix with $H_{i_{jj}} = \left[1 - min\{q, (\frac{\partial\psi_i(\mathbf{Y}_i,A,\mathbf{X}_i;\beta)}{\partial\beta^{\mathrm{T}}} \times \hat{\Gamma})_{jj}\}\right]^{-1/2}$ {Fay and Graubard (2001)}, and $\hat{\psi}_i = \hat{\psi}_i(\mathbf{Y},A,\mathbf{X};\beta)$ is as defined in 1).

3. $\hat{var}_3(\hat{\beta}_{aug}) = \hat{\Gamma}^{-1}\tilde{\Delta}\hat{\Gamma}^{-1^{\mathrm{T}}}$, where $\tilde{\Delta} = m^{-1}\sum_{i=1}^{n}\tilde{\psi}_{i_{opt}}^{\otimes 2}$, and

   $\tilde{\psi}_{i_{opt}} = \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\{\mathbf{Y}_i - \mathbf{g}(A_i;\hat{\beta}_{aug})\} -$
   $(A_i - \pi)\times$
   $\left[\mathbf{D}_i(1)^{\mathrm{T}}\mathbf{V}_i(1)^{-1}\{f_1(\mathbf{X}_i;\hat{\eta}_1) - \mathbf{g}(1;\hat{\beta}_{aug})\} - \mathbf{D}_i(0)^{\mathrm{T}}\mathbf{V}_i(0)^{-1}\{f_0(\mathbf{X}_i;\hat{\eta}_0) - \mathbf{g}(0;\hat{\beta}_{aug})\}\right] -$
   $(A_i - \pi)\times$
   $\left[\mathbf{D}_i(1)^{\mathrm{T}}\mathbf{V}_i(1)^{-1}\{f_1'(\mathbf{X}_i;\hat{\eta}_1)\}\hat{\zeta}_1(\mathbf{Y}_i,\mathbf{X}_i;\hat{\eta}_1) - \mathbf{D}_i(0)^{\mathrm{T}}\mathbf{V}_i(0)^{-1}\{f_0'(\mathbf{X}_i;\hat{\eta}_0)\}\hat{\zeta}_0(\mathbf{Y}_i,\mathbf{X}_i;\hat{\eta}_0)\right].$

   $\hat{\zeta}_k(\mathbf{Y}_i,\mathbf{X}_i;\hat{\eta})$ is the first order approximation of the term $(\hat{\eta}_k - \eta^{*})$ that results from estimation of η_k in $E(\mathbf{Y}_i|\mathbf{X}_i,A_i)$. If $E(\mathbf{Y}_i|\mathbf{X}_i,A_i)$ is estimated by linear regression, $\hat{\zeta}_k(\mathbf{Y}_i,\mathbf{X}_i;\hat{\eta}) = \left(\sum_{i=1}^{n}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{X}_i^{\mathrm{T}}(\mathbf{Y}_i - \mathbf{X}_i\hat{\eta}_k)$. For nonlinear models, in which $E(\mathbf{Y}_i|\mathbf{X}_i,A_i) = \mu(\mathbf{X}_i;\eta_k)$, $(\hat{\eta}_k - \eta_k^{*})$ may be approximated by

$\left(\sum_{i=1}^{n}\mathbf{F}_i^{\mathrm{T}}\mathbf{W}_i^{-1}\mathbf{F}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{F}_i^{\mathrm{T}}\mathbf{W}_i^{-1}(\mathbf{Y}_i-\mu(\mathbf{X}_i;\hat{\eta}_k))$, where $\mathbf{F}_i = \frac{\partial\mu_k(\mathbf{X}_i;\eta)}{\partial\eta}\Big|_{\eta_k=\hat{\eta}_k}$, and $\mathbf{W}_i$ is a diagonal matrix with $W_{i_{jj}}$=$\phi_\mu\nu(\mu)$, following from generalized linear model theory. The function $\nu(\mu)$ denotes the variance function and ϕ_μ the dispersion parameter. We include the subscript μ in ϕ_μ to distinguish from ϕ involved in characterizing $V(Y_{ij}|A_i)$ in the main text.

4. $\hat{var}_4(\hat{\beta}_{aug}) = \hat{\Gamma}^{-1}\tilde{\Delta}^*\hat{\Gamma}^{-1^{\mathrm{T}}}$, where $\tilde{\Delta}^* = m^{-1}\sum_i(\mathbf{H}_i\tilde{\psi}_i)^{\otimes 2}$, with $\tilde{\psi}$ and $\mathbf{H}_i$ are as defined above.

   In 2) and 4), the lower bound q is typically set to 0.75 to prevent gross inflation Fay and Graubard (2001).

Appendix C: Additional Simulations

Table 4.1: **Standard vs. Augmented GEE, Binary Outcome: 250 clusters, low and high association,** $\rho = 0.05$. *Std: unaugmented. Correlation is exchangeable for all estimators. C,F,O,W: augmentation with 'Correct','Forward' selected, 'One-variable', or 'Wrong' model. ML, OLS: augmentation fit with maximum likelihood or ordinary least squares. SE: average unadjusted sandwich. MC RE: square of the Monte Carlo SE of the Std(Exch) estimator divided by the Monte Carlo SE for the indicated estimator. Coverage: coverage based on unadjusted sandwich SE.*

	Estimator	$\hat{\beta}_1$	Bias	SE	MC SE	MC RE	Coverage
m=250, low	Std	-0.3036	0.0077	0.0739	0.0778	1.0000	0.936
	C - ML	-0.3029	0.0069	0.0705	0.0744	1.0951	0.935
	C - OLS	-0.3032	0.0072	0.0710	0.0750	1.0778	0.937
	F - ML	-0.3023	0.0064	0.0701	0.0753	1.0683	0.935
	F - OLS	-0.3026	0.0066	0.0703	0.0757	1.0571	0.937
	O - ML	-0.3033	0.0073	0.0717	0.0752	1.0704	0.937
	O - OLS	-0.3033	0.0073	0.0717	0.0754	1.0658	0.935
	W - ML	-0.3034	0.0075	0.0728	0.0768	1.0271	0.938
	W - OLS	-0.3035	0.0075	0.0728	0.0768	1.0258	0.938
m=250, high	Std	1.1310	0.0052	0.0567	0.0576	1.0000	0.943
	C - ML	1.1314	0.0047	0.0489	0.0497	1.3429	0.940
	C - OLS	1.1313	0.0049	0.0496	0.0505	1.2989	0.943
	F - ML	1.1314	0.0047	0.0485	0.0502	1.3156	0.937
	F - OLS	1.1314	0.0048	0.0491	0.0510	1.2719	0.941
	O - ML	1.1312	0.0050	0.0501	0.0509	1.2799	0.934
	O - OLS	1.1313	0.0048	0.0506	0.0512	1.2651	0.941
	W - ML	1.1310	0.0051	0.0531	0.0542	1.1301	0.938
	W - OLS	1.1310	0.0051	0.0533	0.0542	1.1266	0.937

Table 4.2: **Standard vs. Augmented GEE, Binary Outcome: 250 clusters, low and high association,** $\rho = 0.05$. *Std: unaugmented. Correlation is exchangeable for all estimators. C,F,O,W: augmentation with 'Correct','Forward' selected, 'One-variable', or 'Wrong' model. ML, OLS: augmentation fit with maximum likelihood or ordinary least squares. SE: average unadjusted sandwich. MC RE: square of the Monte Carlo SE of the Std(Exch) estimator divided by the Monte Carlo SE for the indicated estimator. Coverage: coverage based on unadjusted sandwich SE.*

	Estimator	$\hat{\beta}_1$	Bias	SE	MC SE	MC RE	Coverage
m=250, low	Std	-0.2299	0.0135	0.1164	0.1190	1.0000	0.938
	C - ML	-0.2290	0.0126	0.1144	0.1173	1.0293	0.936
	C - OLS	-0.2293	0.0129	0.1146	0.1175	1.0256	0.935
	F - ML	-0.2276	0.0112	0.1135	0.1185	1.0077	0.932
	F - OLS	-0.2280	0.0116	0.1135	0.1187	1.0041	0.933
	O - ML	-0.2294	0.0130	0.1150	0.1175	1.0253	0.935
	O - OLS	-0.2295	0.0130	0.1150	0.1176	1.0234	0.935
	W - ML	-0.2296	0.0131	0.1155	0.1188	1.0020	0.931
	W - OLS	-0.2297	0.0132	0.1155	0.1188	1.0021	0.932
m=250, high	Std	1.0429	0.0072	0.0883	0.0887	1.0000	0.944
	C - ML	1.0436	0.0065	0.0835	0.0848	1.0936	0.949
	C - OLS	1.0435	0.0066	0.0839	0.0851	1.0871	0.948
	F - ML	1.0442	0.0059	0.0828	0.0858	1.0694	0.938
	F - OLS	1.0444	0.0058	0.0831	0.0860	1.0643	0.941
	O - ML	1.0433	0.0068	0.0842	0.0851	1.0863	0.951
	O - OLS	1.0435	0.0067	0.0844	0.0851	1.0861	0.949
	W - ML	1.0431	0.0070	0.0859	0.0869	1.0409	0.951
	W - OLS	1.0431	0.0070	0.0860	0.0869	1.0406	0.950

Appendix D: Deriving the Efficient Score

Let $O_i = (\mathbf{Y}_i, A_i, \mathbf{X}_i)$, where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ...Y_{in_i})^{\mathrm{T}}$ is the n_i-dimensional response vector for the i_{th} independent unit, $i = 1, ..., m$, A_i is a scalar treatment assignment, and $\mathbf{X}_i$ is a matrix of auxiliary covariates. For data O_i augmented estimating functions $\psi_{i_{aug}}(O_i, t; \beta, h, \gamma)$ are constructed by (2.4). The optimal index $h_{opt}(A, t)$ is determined by solving the generalized information equality

$$-E\left[\frac{\partial\psi\{\mathbf{Y}, A, \mathbf{X}, t; \beta, \gamma, h(\cdot)\}}{\partial\beta^{\mathrm{T}}}\bigg|_{\beta=\beta_0}\right] = E\left[\psi\{\mathbf{Y}, A, \mathbf{X}, t; \beta, \gamma, h(\cdot)\}\psi^{\mathrm{T}}\{\mathbf{Y}, A, \mathbf{X}, t; \beta, \gamma, h_{opt}(\cdot)\}\bigg|_{\beta=\beta_0}\right], \tag{4.8}$$

for h_{opt}, where $h(\cdot)$ is any $p \times n_i$ function such that $E[\psi^{\mathrm{T}}\psi] < \infty$.

Conditioning on t, $h(A, t)$ takes up to K different matrix values, $h_0(t), h_1(t), ..., h_{K-1}(t)$, which may be denoted by K $p \times n_i$ constant matrices $\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{K-1}$. Similarly, we define $\Delta_k(\mathbf{X}) = E(\mathbf{Y}|A = k, \mathbf{X}, t) - \mathbf{g}(k, t; \beta)$, the n_i-dimensional vector of the difference in the conditional and marginal mean outcomes under treatment k, where $k = 0, 1, ..., K-1$. Using this construction, let $\mathbf{h} = [\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{K-1}]$ and $\Delta_K(\mathbf{X}) = \{\Delta_0^{\mathrm{T}}(\mathbf{X}), ..., \Delta_{K-1}^{\mathrm{T}}(\mathbf{X})\}^{\mathrm{T}}$. The complete index matrix $\mathbf{h}$ is therefore of dimension $p \times Kn_i$, while Δ_K is a Kn_i-dimensional vector. Estimating functions are then expressed concisely through defining two auxiliary matrix functions of A. Let $\mathbf{A}$ be the $Kn_i \times n_i$ matrix $\mathbf{A} = [I(A = 1)\mathbf{I}_n, ..., I(A = K)\mathbf{I}_n]^{\mathrm{T}}$ and $\mathbf{A}_\pi$ be the $Kn_i \times Kn_i$ block diagonal matrix composed of the diagonal matrices $\{I(A = k) - \pi_k\}\mathbf{I}_n$, where $\mathbf{I}_n$ denotes the $n_i \times n_i$ identity matrix.

Rewriting (2.4) using this notation, we obtain

$$\sum_{i=1}^{n} \mathbf{h}_i\mathbf{A}_i\{\mathbf{Y_i} - \mathbf{g}(A_i, t; \beta)\} - \mathbf{h}\mathbf{A}_{\pi_i i}\Delta_i(\mathbf{X}_i) = \mathbf{0}. \tag{4.9}$$

Substituting this expression into Newey's equations we have

$$E\left[\mathbf{h}\mathbf{A}\frac{\partial \mathbf{g}(A,t;\beta)}{\partial \beta^{\mathrm{T}}}\right] = E\big[\{\mathbf{h}\mathbf{A}(\mathbf{Y}-\mathbf{g}(A,t;\beta)) - \mathbf{h}\mathbf{A}_\pi\Delta_K(\mathbf{X})\}\times \{(\mathbf{Y}-\mathbf{g}(A,t;\beta))^{\mathrm{T}}\mathbf{A}\mathbf{h}_{\mathbf{opt}}^{\mathrm{T}} - \Delta_K^{\mathrm{T}}(\mathbf{X})\mathbf{A}_\pi\mathbf{h}_{\mathbf{opt}}^{\mathrm{T}}\}\big] \tag{4.10}$$

We first note that since $\mathbf{h}$ and $\mathbf{h}_{\mathbf{opt}}$ are constant, we can extract them from the expectation, leaving

$$\mathbf{h}^{\mathrm{T}}E\left[\mathbf{A}\frac{\partial \mathbf{g}(A,t;\beta)}{\partial \beta^{\mathrm{T}}}\right] = \mathbf{h}^{\mathrm{T}}E\big[\{\mathbf{A}(\mathbf{Y}-\mathbf{g}(A,t;\beta)) - \mathbf{A}_\pi\Delta_K(\mathbf{X})\}\times \{(\mathbf{Y}-\mathbf{g}(A,t;\beta))^{\mathrm{T}}\mathbf{A} - \Delta_K^{\mathrm{T}}(\mathbf{X})\mathbf{A}_\pi\}\big]\mathbf{h}_{\mathbf{opt}}^{\mathrm{T}} \tag{4.11}$$

Since $\mathbf{h}$ is nonzero, it must hold that

$$E\left[\mathbf{A}\frac{\partial \mathbf{g}(A,t;\beta)}{\partial \beta^{\mathrm{T}}}\right] = E\left[\{\mathbf{A}(\mathbf{Y}-\mathbf{g}(A,t;\beta)) - \mathbf{A}_\pi\Delta_K(\mathbf{X})\}\{(\mathbf{Y}-\mathbf{g}(A,t;\beta))^{\mathrm{T}}\mathbf{A} - \Delta_K^{\mathrm{T}}(\mathbf{X})\mathbf{A}_\pi\}\right]\mathbf{h}_{\mathbf{opt}}^{\mathrm{T}} \tag{4.12}$$

Evaluating the left hand side of the equation, we have

$$E\left\{\begin{bmatrix} A_0\mathbf{I}_n \\ A_2\mathbf{I}_n \\ \vdots \\ A_{K-1}\mathbf{I}_n \end{bmatrix}\frac{\partial \mathbf{g}(A,t;\beta)}{\partial \beta^{\mathrm{T}}}\right\} = \begin{bmatrix} \pi_0\frac{\partial \mathbf{g}(0,t;\beta)}{\partial \beta^{\mathrm{T}}} \\ \pi_1\frac{\partial \mathbf{g}(1,t;\beta)}{\partial \beta^{\mathrm{T}}} \\ \vdots \\ \pi_{K-1}\frac{\partial \mathbf{g}(K-1,t;\beta)}{\partial \beta^{\mathrm{T}}} \end{bmatrix} \tag{$\mathbf{D}^*$}$$

Evaluating the right hand side, we note that we have an expression of the form $E[A-B][A_{opt}-B_{opt}]^{\mathrm{T}}$. Interpreting the augmented estimating function as a residual, we note that $A - B \perp B_{opt}$. We can therefore evaluate $E[A-B][A_{opt}-B_{opt}]^{\mathrm{T}} = E[A-B][A_{opt}]^{\mathrm{T}}$. In (4.12), this becomes

$$E[\mathbf{A}\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}\{\mathbf{Y}-g(A,t;\beta)\}^{\mathrm{T}}\mathbf{A}] - E[\mathbf{A}_\pi\Delta(\mathbf{X})\{\mathbf{Y}-g(A,t;\beta)\}^{\mathrm{T}}\mathbf{A}] \tag{4.13}$$

Regarding the first term in (4.13), we have

$$E[\mathbf{A}\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}\{\mathbf{Y}-g(A,t;\beta)^{\mathrm{T}}\}\mathbf{A}] = E\begin{bmatrix} A_0A_0\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} & \cdots & A_0A_{K-1}\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} \\ A_1A_0\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} & \cdots & A_1A_{K-1}\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} \\ \vdots & \ddots & \vdots \\ A_{K-1}A_0\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} & \cdots & A_{K-1}A_{K-1}\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} \end{bmatrix}, \tag{4.14}$$

where $U^{\otimes 2}$=UU^{T}. Since each individual is only assigned to one treatment, only one of A_0, A_1,...,A_{K-1} is nonzero. The non diagonal blocks of (4.14) are identically 0. The diagonal blocks contain terms of the form $E[A_k A_k\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2}] = E[A_k\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2}] = \pi_k V(\mathbf{Y}|A=k)$.

Matrix (4.14) is written as

$$\begin{bmatrix} \pi_0 V(\mathbf{Y}|A=0) & 0 & \cdots & 0 \\ 0 & \pi_1 V(\mathbf{Y}|A=1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_{K-1} V(\mathbf{Y}|A=K-1) \end{bmatrix} \qquad (\mathbf{C}_1)$$

Evaluating the second term of (4.13), we have

$$\begin{aligned} &E[\mathbf{A}_\pi \Delta_K(\mathbf{X})\{\mathbf{Y}-g(A,t,\beta)\}^{\mathrm{T}}\mathbf{A}] = \\ &E\Bigg\{ \begin{bmatrix} (A_0-\pi_0)\mathbf{I}_n & \ldots & 0 \\ \vdots & (A_1-\pi_1)\mathbf{I}_n & \vdots \\ 0 & \cdots & (A_{K-1}-\pi_{K-1})\mathbf{I}_n \end{bmatrix} \begin{bmatrix} \Delta_0(\mathbf{X}) \\ \Delta_1(\mathbf{X}) \\ \vdots \\ \Delta_{K-1}(\mathbf{X}) \end{bmatrix} \times \\ &\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\mathrm{T}} \begin{bmatrix} A_0\mathbf{I}_n & A_1\mathbf{I}_n & \cdots & A_{K-1}\mathbf{I}_n \end{bmatrix} \Bigg\} \\ =&E\left\{ \begin{bmatrix} (A_0-\pi_0)A_0\Delta_0(\mathbf{X})\Delta_A^{\mathrm{T}}(\mathbf{X}) & \cdots & (A_0-\pi_0)A_{K-1}\Delta_0(\mathbf{X})\Delta_A^{\mathrm{T}}(\mathbf{X}) \\ (A_1-\pi_1)A_0\Delta_1(\mathbf{X})\Delta_A^{\mathrm{T}}(\mathbf{X}) & \ddots & (A_1-\pi_1)A_{K-1}\Delta_1(\mathbf{X})\Delta_A^{\mathrm{T}}(\mathbf{X}) \\ \vdots & \ddots & \vdots \\ (A_{K-1}-\pi_{K-1})A_0\Delta_{K-1}(\mathbf{X})\Delta_A^{\mathrm{T}}(\mathbf{X}) & \cdots & (A_{K-1}-\pi_{K-1})A_{K-1}\Delta_A^{\mathrm{T}}(\mathbf{X}) \end{bmatrix} \right\} \\ =&\begin{bmatrix} \pi_0(1-\pi_0)\Delta_0(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X}) & \cdots & -\pi_0\pi_{K-1}\Delta_0(\mathbf{X})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X}) \\ -\pi_1\pi_0\Delta_1(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X}) & \ddots & -\pi_1\pi_{K-1}\Delta_1(\mathbf{X})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X}) \\ \vdots & \ddots & \vdots \\ -\pi_{K-1}\pi_0\Delta_{K-1}(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X}) & \cdots & \pi_{K-1}(1-\pi_{K-1})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X}) \end{bmatrix} \end{aligned} \qquad (\mathbf{C_2})$$

From ($\mathbf{C_2}$), we see that generally, the second term of (4.13) contains block diagonal terms $\pi_k(1-\pi_k)E_{\mathbf{X}}\left\{\Delta_k^{\otimes 2}(\mathbf{X})\right\}$, and block off-diagonal terms $-\pi_j\pi_k E_{\mathbf{X}}\{\Delta_j(\mathbf{X})\Delta_k^{\mathrm{T}}(\mathbf{X})\}$.

Referring to (4.12), we see that $\mathbf{h_{opt}} = [\mathbf{C}_1 - \mathbf{C}_2]^{-1}\mathbf{D}^*$, as labeled above.

References

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press, Baltimore.

BICKEL, P. J. and ZWET, W. R. V. (1978). Asymptotic expansions for the power of distribution-free tests in the two-sample problem. *The Annals of Statistics* **6** 937–1004.

BLACK, P. (2005). Greedy algorithm. In *Dictionary of Algorithms and Data Structures* (P. Black, ed.).
URL `http://xlinux.nist.gov/dads//HTML/greedyalgo.html`

BRAUN, T. M. and FENG, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *Journal of the American Statistical Association* **96** 1424–1432. DOI: 10.1198/016214501753382336.

CHAMBERLAIN, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics* **32** 189–218.

DAVIDSON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press.

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *The Annals of Statistics* **32** 407–499.

FAN, J. and LI, R. (2001). Variable selection via nonconvace penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

FAY, M. P. and GRAUBARD, B. I. (2001). Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics* **57** 1198–1206. DOI: 10.1111/j.0006-341X.2001.01198.x.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd.

GAIL, M. H., MARK, S. D., CARROLL, R. J., GREEN, S. B. and PEE, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* **15** 1069–1092. DOI: 0.1002/(SICI)1097-0258(19960615)15:11¡1069::AID-SIM220¿3.3.CO;2-H.

GAIL, M. H., TAN, W. Y. and PIANTADOSI, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika* **75** 57–64.

GUNSOLLEY, J. C., GETCHELL, C. and CHINCHILLI, V. M. (1995). Small sample characteristics of generalized estimating equations. *Communications in Statistics - Simulation and Computation* **24** 869–878. DOI: 10.1080/03610919508813280.

HAMMER, S. M., VAIDA, F., BENNETT, K., HOLOHAN, M. K., SHEINER, L., ERON, J., WHEAT, L. J., MITSUYASU, R. T., GULICK, R. M., VALENTINE, F. T., ABERG, J. A., ROGERS, M. D., KAROL, C. N., SAAH, A. J., LEWIS, R. H., BESSEN, L. J., BROSGART, C., DE GRUTTOLA, V. and MELLORS, J. W. (2002). Dual vs. single protease inhibitor therapy following antiretroviral treatment failure. *Journal of the American Medical Association* **288** 169–180.

HUBER, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35** 73–101.

KAMO, N., CARLSON, M., BRENNAN, R. T. and EARLS, F. (2008). Young citizens as health agents: Use of drama in promoting community efficacy for hiv/aids. *American Journal of Public Health* **98** 201–204. DOI: 10.2105/AJPH.2007.113704.

KAUERMANN, G. and CARROLL, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96** 1387–1396. DOI: 10.1198/016214501753382309.

KLAR, N. and DONNER, A. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Hodder Arnold.

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974. DOI: 10.2307/2529876.

LEON, A., TSIATIS, A. A. and DAVIDIAN, M. (2003). Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics* **59** 1046–1055. DOI: 10.1111/j.0006-341X.2003.00120.x.

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130. DOI: 10.2307/2531248.

LIPSITZ, S. R., FITZMAURICE, G. M., ORAV, E. J. and LAIRD, N. M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics* **50** 270–278. DOI: 10.2307/2533218.

MANCL, L. A. and DEROUEN, T. A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics* **57** 126–134. DOI: 10.1111/j.0006-341X.2001.00126.x.

MOORE, K. L. and VAN DER LAAN, M. J. (2009a). Application of time-to-event methods in the assessment of safety in clinical trials. In *Design, Summarization, Analysis & Interpretation of Clinical Trials with Time-to-Event Endpoints* (K. E. Peace, ed.). Chapman & Hall.

MOORE, K. L. and VAN DER LAAN, M. J. (2009b). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine* **28** 39–64. DOI: 10.1002/sim.3445.

MURRAY, D. M., VARNELL, S. P. and BLITSTEIN, J. L. (2004). Design and analysis of group randomized trials. *American Journal of Public Health* **94** 423–432. DOI: 10.2105/AJPH.94.3.423.

NEWEY, W. K. and MCFADDEN, D. (1994). Chapter 36 large sample estimation and hypothesis testing. vol. 4 of *Handbook of Econometrics*. Elsevier, 2111 – 2245. URL http://www.sciencedirect.com/science/article/pii/S1573441205800054

PAN, W. and WALL, M. M. (2002). Small-sample adjustments in using the sandwich estimator in generalized estimating equations. *Statistics in Medicine* **21** 1429–1441. DOI: 10.1002/sim.1142.

PAULER, D. K. (1998). The schwarz criterion and related methods for normal linear models. *Biometrika* **85** 13–27.

POCOCK, S. J., ASSMAN, S. E., ENOS, L. E. and KASTEN, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* **21** 2917–2930. DOI: 10.1002/sim.1296.

ROBINS, J. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999*.

ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modelling* **7** 1393–1512.

ROBINS, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials* (D. Berry and M. E. Halloran, eds.), vol. 116. NY: Springer-Verlag, 95–134.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89** 846–866. DOI: 10.2307/2290910.

ROBINSON, L. D. and JEWELL, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* **58** 227–240.

ROSENBAUM, P. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science* **17** 286–327.

RUBIN, D. and VAN DER LAAN, M. J. (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics* **4**. DOI: 10.2202/1557-4679.1084.

SOFER, T., DICKER, L. and LIN, X. (2012). Variable selection for high-dimensional multivariate outcomes. In preparation.

STEPHENS, A. J., TCHETGEN TCHETGEN, E. J. and DE GRUTTOLA, V. (2012a). Augmented gee for improving efficiency of inferences in cluster randomized trials by leveraging cluster and individual-level covariates. *Statistics in Medicine* In press.

STEPHENS, A. J., TCHETGEN TCHETGEN, E. J. and DE GRUTTOLA, V. (2012b). Locally efficient estimation of marginal treatment effects using auxiliary covariates in randomized trials with correlated outcomes. In preparation.

THORNQUIST, A. and ANDERSON, G. L. (1992). Small-sample properties of generalized estimating equations in group randomized designs with gaussian response Retrieved from author.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58** 267–288.

TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LU, X. (2008). Covariate adjustment for two-sample treatment comparisons for randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* **27** 4658–4677. DOI: 10.1002/sim.3113.

TUGLUS, C. and VAN DER LAAN, M. J. (2010). Targeted maximum likelihood method for repeated measures semiparametric regression: Discovery for transcription factor activity. *U.C. Berkeley Division of Biostatistics Working Paper Series* .

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. NY: Springer-Verlag.

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2** 1–40. DOI: 10.2202/1557-4679.1043.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

WANG, A. and LOUIS, T. A. (2003). Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika* **90** 765–775. DOI: 10.1093/biomet/90.4.765,.

WANG, Y. and CAREY, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. *Biometrika* **90** 29–41.

ZHANG, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Journal of the American Statistical Association* **101** 1418–1429.

ZHANG, M. and GILBERT, P. B. (2010). Increasing the efficiency of prevention trials by incorporating baseline covariates. *Statistical Communications in Infectious Disease* **2**. DOI: 10.2202/1948-4690.1002.

ZHANG, M., TSIATIS, A. A. and DAVIDIAN, M. (2008). Improving efficiency of inferences in clinical randomized trials using auxiliary covariates. *Biometrics* **64** 707–715. DOI: 10.1111/j.1541-0420.2007.00976.x.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.